KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Science

Extreme Value Theory in Finance and Insurance

Tom Reynkens

Supervisors:
Prof. dr. J. Beirlant
Prof. dr. W. Schoutens
Prof. dr. T. Verdonck

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Science (PhD): Mathematics

June 2017

Extreme Value Theory in Finance and Insurance

Tom REYNKENS

Examination committee:
Prof. dr. A. Carbonez, chair
Prof. dr. J. Beirlant, supervisor
Prof. dr. W. Schoutens, co-supervisor
Prof. dr. T. Verdonck, co-supervisor
Prof. dr. K. Antonio
Dr. J. De Spiegeleer
Prof. dr. J. Dhaene
Prof. dr. H. Albrecher
(Université de Lausanne, Switzerland)

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Science (PhD): Mathematics

June 2017

Uitgegeven in eigen beheer, Tom Reynkens, Celestijnenlaan 200B, B-3001 Leuven (Belgium)

Acknowledgements

When I started my studies back in 2008, the field of statistics was relatively unknown to me. I never imagined that I would write a master's thesis in that branch of applied mathematics, and I could certainly not have imagined that nine years later I would finish a PhD thesis in statistics. Looking back on the previous four years, I noticed that many people played an important role in the realisation of this thesis, and I want to thank all of them.

I sincerely want to thank Jan Beirlant for the three years of collaborations which led to many interesting research projects, exciting connections with the industry and unexpected new academic roads. The countless discussions, in real life and over the phone, were always pleasant and provided many new insights. *Baie dankie* for all opportunities which allowed me to learn a lot and meet very interesting people.

It took some time and effort to convince me to start a PhD, but I am very glad that Tim Verdonck persisted. I enjoyed our many collaborations both in research and education, and will always remember the comical remarks during our meetings. I also want to thank Wim Schoutens for giving me the opportunity to start a PhD and introducing me to the world of financial research.

I am very grateful to my jury for taking the time to follow up on my PhD progress and to assess this dissertation. Their remarks and suggestions over the years were very helpful for my research. I very much appreciate the work of An Carbonez as the chair of my PhD jury which made the defences go smoothly.

Academic research is a team effort and therefore I would like to thank all my co-authors for the nice collaborations and discussions on several papers. Special thanks go to Roel Verbelen and Eric Schmitt for the long meetings where a lot of crucial ideas originated. I really enjoyed working together! I would also like to thank Hansjörg Albrecher for the interesting discussions on the reinsurance book which greatly helped improve my research, and for the nice collaborations during the workshops all over the world.

During the four years of my PhD, fellow members of the section made life really enjoyable. Special thanks go to Florence, Monika, Ine, Jakob, Sebastiaan, Pieter and Kris, for always being available for a nice chat, jokes or a coffee break when it was needed. I would also like to thank my colleagues for the nice lunch breaks in the "staff lounge" (a.k.a. "aquarium"), even though they could not always convince me to play cards, and for the inter-sectional football games. Last but not least, a big thank you to the colleagues from ORSTAT and AFI for the fun after-work activities such as the dinners, various sports, almost winning the EDC quiz, and hence organising it the next year, and supporting the Red Devils.

Graag wil ik ook mijn vrienden bedanken voor de vele leuke momenten zoals de etentjes, de weekends weg en samen Nieuwjaar vieren. Speciale momenten tijdens het academiejaar waren zeker de quizzen met "De Wiskies" die vaak voor spannende momenten zorgden.

Deze thesis was niet mogelijk geweest zonder de steun van mijn familie en schoonfamilie. Bedankt om er altijd te zijn voor mij. Graag wil ik mijn mama bedanken voor de goede zorgen en het vele heerlijke eten.

Tenslotte wil ik Liese bedanken om mijn grote steun te zijn tijdens mijn doctoraat. Het was soms nogal druk waardoor ik niet veel tijd voor je had, maar je was altijd geduldig en heel erg behulpzaam!

Tom Reynkens
Leuven & Heusden-Zolder
June 2017

Abstract

When modelling high-dimensional data, dimension reduction techniques such as principal component analysis (PCA) are often used. In the first part of this thesis we will focus on two drawbacks of classical PCA. First, interpretation of classical PCA is often challenging because most of the loadings are neither very small nor very large in absolute value. Second, classical PCA can be heavily distorted by outliers since it is based on the classical covariance matrix. In order to resolve both problems, we present a new PCA algorithm that is robust against outliers and yields sparse PCs, i.e. PCs with many zero loadings. The approach is based on the ROBPCA algorithm that generates robust but non-sparse loadings. The construction of the new ROSPCA method is detailed, as well as a selection criterion for the sparsity parameter. An extensive simulation study and a real data example are performed, showing that it is capable of accurately finding the sparse structure of datasets, even when challenging outliers are present.

Stock market crashes such as Black Monday in 1987 and catastrophes such as earthquakes are examples of extreme events in finance and insurance, respectively. They are large events with a considerable impact that occur seldom. Extreme value theory (EVT) provides a theoretical framework to model extreme values such that e.g. risk measures can be estimated based on available data. In the second part of this PhD thesis we focus on applications of EVT that are of interest to finance and insurance.

A Black Swan is an improbable event with massive consequences. We propose a way to investigate if the 2007–2008 financial crisis was a Black Swan event for a given bank based on weekly log-returns. This is done by comparing the tail behaviour of the negative log-returns before and after the crisis using techniques from extreme value methodology. We illustrate this approach with Barclays and Credit Suisse data, and then link the differences in tail risk behaviour between these banks with economic indicators.

The earthquake engineering community, disaster management agencies and the insurance industry need models for earthquake magnitudes to predict possible damage by earthquakes. A crucial element in these models is the area-characteristic, maximum possible earthquake magnitude. The Gutenberg-Richter distribution, which is a (doubly) truncated exponential distribution, is widely used to model earthquake magnitudes. Recently, Aban et al. (2006) and Beirlant et al. (2016a) discussed tail fitting for truncated Pareto-type distributions. However, as is the case for the Gutenberg-Richter distribution, in some applications the underlying distribution appears to have a lighter tail than the Pareto distribution. We generalise the classical peaks over threshold (POT) approach to allow for truncation effects. This enables a unified treatment of extreme value analysis for truncated heavy and light tails. We use a pseudo maximum likelihood approach to estimate the model parameters and consider extreme quantile estimation. The new approach is illustrated on examples from hydrology and geophysics. Moreover, we perform simulations to illustrate the potential of the method on truncated heavy and light tails.

The new approach can then be used to estimate the maximum possible earthquake magnitude. We also look at two other EVT-based endpoint estimators and endpoint estimators that are used in the geophysical literature. To quantify uncertainty of the point estimates for the endpoint, upper confidence bounds are also considered. We apply the techniques to provide estimates, and upper confidence bounds, for the maximum possible earthquake magnitude in Groningen where earthquakes are induced by gas extraction. Furthermore, we compare the methods from extreme value theory and the geophysical literature through simulations.

In risk analysis, a global fit that appropriately captures the body and the tail of the distribution of losses is essential. Modelling the whole range of the losses using a standard distribution is usually very hard and often impossible due to the specific characteristics of the body and the tail of the loss distribution. A possible solution is to combine two distributions in a splicing model: a light-tailed distribution for the body which covers light and moderate losses, and a heavy-tailed distribution for the tail to capture large losses. We propose a splicing model with the flexible mixed Erlang distribution for the body and a Pareto distribution for the tail. Motivated by examples in financial risk analysis, we extend our splicing approach to censored and/or truncated data. We illustrate the flexibility of this splicing model using practical examples from reinsurance.

Beknopte samenvatting

Om data met veel dimensies te modelleren worden vaak dimensiereductie-technieken zoals hoofdcomponentenanalyse (PCA) gebruikt. In het eerste deel van de thesis focussen we op twee nadelen van klassieke PCA. Ten eerste is de interpretatie van klassieke PCA vaak moeilijk omdat veel van de componenten van de PCA-richtingen niet heel groot of heel klein zijn (in absolute waarde). Ten tweede is klassieke PCA gevoelig aan uitschieters omdat het gebaseerd is op de klassieke covariantiematrix. Om beide problemen op te lossen, stellen we een nieuw PCA-algoritme voor dat robuust is tegen uitschieters en schaarse hoofdcomponenten geeft, d.w.z. PCA-richtingen met veel componenten die nul zijn. De methode is gebaseerd op het ROBPCA algoritme dat robuuste maar niet-schaarse PCA-richtingen genereert. We bespreken zowel de constructie van dit algoritme als een selectiecriterium voor de schaarsheidsparameter. Op basis van een uitgebreide simulatiestudie en een datavoorbeeld kunnen we besluiten dat ROSPCA de schaarse structuur van datasets kan vinden, zelfs als uitdagende uitschieters aanwezig zijn.

Beurscrashes zoals Zwarte Maandag in 1987 en rampen zoals aardbevingen zijn respectievelijk voorbeelden van extreme gebeurtenissen in financiën en verzekeringen. Deze gebeurtenissen hebben een grote impact en zijn redelijk zeldzaam. Extreme waarde theorie (EVT) geeft een theoretisch kader om extreme waarden te modelleren zodat bv. risicomaten geschat kunnen worden a.d.h.v. de beschikbare data. In het tweede deel van deze thesis focussen we op toepassingen van EVT die interessant zijn voor financiën en verzekeringen.

Een gebeurtenis die onwaarschijnlijk geacht wordt maar wel gigantische gevolgen heeft, wordt in de financiële wereld een Zwarte Zwaan genoemd. We stellen een manier voor om te onderzoeken, op basis van wekelijkse log-rendementen, of de financiële crisis van 2007–2008 een Zwarte Zwaan gebeurtenis is voor een bank. We doen dit door het staartgedrag van de negatieve log-rendementen voor en na de crisis te onderzoeken met EVT technieken. We illustreren deze methode bij twee bekende Europese banken: Barclays en Credit Suisse, en bespreken

daarna de verschillen in risicogedrag a.d.h.v. economische indicatoren.

Om schade veroorzaakt door aardbevingen te voorspellen, hebben aardbevingsingenieurs, crisiscentra en verzekeringsbedrijven modellen nodig voor de magnitudes van aardbevingen. Een cruciale grootheid in deze modellen is de maximaal mogelijke aardbevingsmagnitude. Een veelgebruikt model voor magnitudes is de Gutenberg-Richter verdeling: een (dubbel) afgeknotte exponentiële verdeling. Recent hebben Aban e.a. (2006) en Beirlant e.a. (2016a) modellen voor de staart van afgeknotte Pareto-achtige verdelingen bestudeerd. Maar in sommige toepassingen, zoals bij de Gutenberg-Richter verdeling, lijkt de onderliggende verdeling een lichtere staart te hebben dan de Pareto verdeling. Daarom veralgemenen we de peaks over threshold (POT) techniek naar afknotte verdelingen. Dit zorgt voor een algemene aanpak van extreme waarde analyses voor afgeknotte zwaarstaartige en afgeknotte lichtstaartige verdelingen. We gebruiken een meest-aannemelijkheids-aanpak om de parameters van het model te schatten, en we bestuderen ook de schatting van extreme kwantielen. De nieuwe aanpak wordt geïllustreerd op voorbeelden uit hydrologie en geofysica. Daarnaast voeren we ook simulaties uit om het potentieel van de methode te illustreren op afgeknotte zware en lichte staarten.

Met de nieuwe methode kunnen we dan de maximaal mogelijke aardbevingsmagnitude schatten. We bekijken bovendien ook twee andere eindpunt schatters op basis van EVT, en schatters uit de geofysische literatuur. Om de onzekerheid van de puntschatters voor het eindpunt te kwantificeren, bestuderen we ook de bovengrens van het betrouwbaarheidsinterval voor deze parameter. We gebruiken de methodes dan om de maximaal mogelijke aardbevingsmagnitude in Groningen te schatten waar aardbevingen geïnduceerd worden door gaswinning. Bovendien vergelijken we de methodes op basis van EVT en uit de geofysische literatuur via simulaties.

In risico-analyses is het cruciaal om een globaal model te hebben dat het gedrag van zowel de hoofdmoot als de staart van de verdeling van de verliezen beschrijft. Alle verliezen tegelijk modelleren met een standaardverdeling is meestal heel erg moeilijk of zelfs onmogelijk door de specifieke eigenschappen van de hoofdmoot en de staart van de verdeling van de verliezen. Een mogelijke oplossing is om beide delen apart te modelleren en ze dan te combineren in een verbindingsmodel: een lichtstaartige verdeling voor de kleine en de middelgrote verliezen, en een zwaarstaartige verdeling voor de grote verliezen. We stellen een verbindingsmodel voor met de flexibele gemixte Erlang verdeling voor de hoofdmoot van de verdeling, en een EVT-verdeling om de staart te modelleren. Bovendien breiden we de verbindingsaanpak uit naar gecensureerde en afgeknotte data aangezien dit soort data vaak voorkomt bij financiële risico-analyses. We illustreren tenslotte de flexibiliteit van het verbindingsmodel met herverzekeringsvoorbeelden.

List of abbreviations

AD	Anderson-Darling
AIC	Akaike Information Criterion
AO	Adjusted Outlyingness
BIC	Bayesian Information Criterion
CDF	Cumulative Distribution Function
CLT	Central Limit Theorem
CPCA	Classical Principal Component Analysis
CPV	Cumulative Percent Variation
EM	Expectation-Maximisation
EPD	Extended Pareto Distribution
EPXMA	Electron Probe X-ray Microanalysis
EVA	Extreme Value Analysis
EVI	Extreme Value Index
EVT	Extreme Value Theory
GARCH	Generalised Autoregressive Conditional Heteroskedasticity
GEV	Generalised Extreme Value
GoF	Goodness-of-Fit
GPD	Generalised Pareto Distribution
GR	Gutenberg-Richter
IC	Information Criterion
KS	Kolmogorov-Smirnov

LASSO	Least Absolute Shrinkage and Selection Operator
LOB	Line of Business
MCD	Minimum Covariance Determinant
MDA	Max-Domain of Attraction
ME	Mixed Erlang
MGPD	Multivariate Generalised Pareto Distribution
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MSE	Mean Squared Error
MTPL	Motor Third Party Liability
NLL	Negative Log-Likelihood
OD	Orthogonal Distance
PC	Principal Component
PCA	Principal Component Analysis
PDF	Probability Density Function
POT	Peaks Over Threshold
PP	Probability-Probability
PP-PCA	Projection Pursuit Principal Component Analysis
QQ	Quantile-Quantile
ROSCPA	RObust Sparse Principal Component Analysis
RSS	Residual Sum of Squares
RTF	Right Tail Function
SCoTLASS	Simplified Component Technique - LASSO
SD	Score Distance
SVD	Singular Value Decomposition
TVaR	Tail Value-at-Risk
VaR	Value-at-Risk

Contents

Chapter 1

Introduction

In the first part of this PhD thesis, we introduce a robust sparse principal component analysis method. In the second part we develop several methods in extreme value theory with applications in finance and insurance. This chapter contains an introduction to both parts.

1.1 Part I: Robust Sparse PCA

When modelling high-dimensional data, dimension reduction techniques are often used to ease interpretation. One of the most popular dimension reduction techniques is principal component analysis (PCA) which was developed independently by Karl Pearson (1901) and Harold Hotelling (1933; 1936) in the first part of the 20th century. The idea of PCA is to find a transformation of the variables such that the transformed variables are uncorrelated and capture most of the covariance structure of the original data. The transformed variables, which are linear combinations of the original variables, are called the principal components (PCs). Those directions are thus chosen such that they are orthogonal and sequentially maximise the variance of the projected data. The classical PCA directions correspond to the eigenvectors of the sample covariance matrix, and the variance of the data projected on an eigenvector is equal to the corresponding eigenvalue. To reduce dimensions, not all the PCs are used, but only a limited set of them which explain an adequate amount of the total variance of the original data. A detailed overview of PCA and its properties can be found in Jolliffe (2002).

PCA is widely used in many fields such as chemometrics, economics, quality control and signal processing. In finance, PCA is for example used for interest rate modelling, see Pelata et al. (2012) and Ruppert (2010). Another interesting application of PCA can be found in fraud detection. RIDIT scores (Bross, 1958) are used to quantify the level of fraud suspicion for an individual/object based on ordered categorical variables. Brockett et al. (2002) proposed the PRIDIT method which uses the first PC of RIDIT scores for different variables. This gives an overall fraud score for each individual/object which is a weighted sum of the RIDIT scores. These weights are thus chosen to maximise the captured variability of the RIDIT scores. Brockett et al. (2002) applied the method to detect fraud for bodily injury claims in automobile insurance.

As an example we consider the Kibler car dataset (Kibler et al., 1989) which contains 14 variables that describe 195 car models. Four PCs explain around 83% of the total variance. The PC loadings, i.e. the components of the linear combinations, are given in Table 1.1. The first transformed variable is given by $0.11\ \textit{symboling} + \ldots + (-0.32)\ \textit{price}$, and it describes around 50% of the total variance. Instead of working with the 14 original variables, we look at 4 new variables in the PCA subspace.

	PC1	PC2	PC3	PC4
symboling	0.11	0.38	-0.37	-0.32
wheel-base	-0.32	-0.27	0.08	0.16
length	-0.35	-0.13	0.05	0.05
width	-0.34	-0.08	-0.12	0.04
height	-0.14	-0.40	0.39	0.17
curb-weight	-0.37	-0.02	-0.11	-0.04
bore	-0.27	0.03	0.07	-0.41
stroke	-0.05	-0.08	-0.64	0.59
compression-ratio	-0.02	-0.45	-0.41	-0.20
horsepower	-0.30	0.32	-0.12	-0.02
peak-rpm	0.09	0.39	0.20	0.52
city-mpg	0.32	-0.28	-0.10	-0.05
highway-mpg	0.33	-0.23	-0.10	-0.05
price	-0.32	0.11	-0.15	-0.07

Table 1.1: Kibler dataset: PCA loadings.

In this dissertation we will focus on two drawbacks of classical principal component analysis (CPCA): difficult interpretation of the loadings and influence of outliers on the estimates.

Interpretation of classical PCs is often difficult because most of the loadings are neither very small nor very large in absolute value. Therefore, sparse PCA methods were developed to estimate PCs with many zero loadings which increases interpretability. Two widely used sparse PCA methods are SCoTLASS (Jolliffe et al., 2003) and SPCA (Zou et al., 2006).

Applying SCoTLASS to the Kibler dataset results in the loadings matrix in Table 1.2. A lot of the loadings are now estimated to be zero which eases interpretation. We see that the second and third PC are determined by only two variables each, and that the fourth PC is equal to a single variable: "stroke".

	PC1	PC2	PC3	PC4
symboling	0	0.71	0	0
wheel-base	0.29	0	0	0
length	0.36	0	0	0
width	0.35	0	0	0
height	0	-0.71	0	0
curb-weight	0.39	0	0	0
bore	0.25	0	0	0
stroke	0	0	0	1
compression-ratio	0	0	-0.71	0
horsepower	0.32	0	0	0
peak-rpm	0	0	0.71	0
city-mpg	-0.34	0	0	0
highway-mpg	-0.36	0	0	0
price	0.34	0	0	0

Table 1.2: Kibler dataset: PCA loadings obtained using SCoTLASS.

When applying classical statistical methods, it is assumed that the data come from a specified distribution. In practice, this assumption is frequently violated. However, when the majority of the data do come from this distribution, robust methods can be applied. The observations that deviate from the majority are then called outliers. Robust methods fit the majority of the data well, if not too many outliers are present, and yield approximately the same results as the classical methods when there are no outliers. Because they provide an appropriate fit for the bulk of the data, robust methods can be used to detect influential data points that need further investigation. This is our main goal when applying robust methods to financial and actuarial data.

Robust methods have been around for at least 200 years, but major improvements have been made since the end of the 1960s, based on work by John Tukey (1960), Peter Huber (1964; 1967) and Frank Hampel (1971; 1974). Fundamental books on robust statistics have been written in the next 20 years including Huber (1981) and Hampel et al. (1986). Robust alternatives have been developed for many (classical) statistical procedures such as estimation of the covariance matrix, least squares regression, and time series analysis. An overview of robust methods and their applications can be found in Maronna et al. (2006).

Since CPCA is based on the classical covariance matrix, it can be heavily distorted by outliers. Therefore, several robust alternatives for CPCA have been proposed including a projection pursuit principal component analysis (PP-PCA)

approach (Li and Chen, 1985; Hubert et al., 2002; Croux and Ruiz-Gazen, 2005), spherical PCA (Locantore et al., 1999), PCA using a robust M-scale estimator (Maronna, 2005), and ROBPCA (Hubert et al., 2005). The SCoTLASS and SPCA methods are also heavily influenced by outliers, and can hence yield unreliable estimates in the presence of outliers.

ROBPCA can be employed to robustly estimate the PCA subspace for the Kibler dataset. Using outlier detection techniques based on this subspace, we find 20 deviating observations. Closer inspection reveals that they correspond to the 20 diesel cars in the data, whereas all other cars in the dataset use petrol. Only one of the diesel cars is detected using the classical PCA subspace since it is influenced by these outliers. This effect is also present for other non-robust estimators and is known as masking. Similarly, none of the diesel cars is found to be outlying based on the sparse PCA subspace estimated by SCoTLASS since it is not robust.

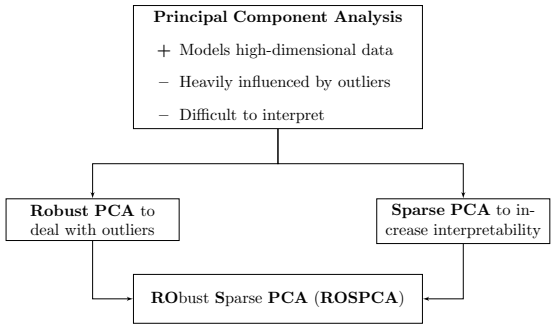


Figure 1.1: Overview of robust sparse PCA.

In order to resolve both problems, we present a new sparse PCA algorithm which is robust against outliers, see Figure 1.1. The approach is based on the ROBPCA algorithm that generates robust but non-sparse loadings. In Chapter 2, the construction of the new ROSPCA method is detailed, as well as a selection criterion for the sparsity parameter. An extensive simulation study and a real data example are performed, showing that it is capable of accurately finding the sparse structure of datasets, even when challenging outliers are present. Previous work on this problem can be found in Croux et al. (2013). In comparison with their projection pursuit-based algorithm, ROSPCA demonstrates improved robustness properties and sparsity estimation capability, as well as significantly faster computation time. Moreover, we propose an adjusted version of ROSPCA that can handle skewed data and apply it to a financial data example.

1.2 Part II: Extreme Value Theory in Finance and Insurance

What height should a dyke be such that it can withstand a once-in-10 000-years storm? This simple question is crucial for the Netherlands since a large part of the country is below the sea level and dykes protect it from flooding. Statisticians and engineers try to answer this question based on flood data which leads to an important problem: how to estimate the size of such a storm when only around 100 years of data is available? This is a typical example of extreme event analysis. Embrechts et al. (1997) give two properties of extreme events: they are large events with a considerable impact and they are rare events. In probability terms they can be defined as events in the right tail of the distribution. Because of their big impact it is crucial to model them, but since few of these large observations are available, specific modelling techniques are needed. Extreme value theory (EVT) provides a theoretical framework to model extreme values such that e.g. small exceedance probabilities or large quantiles can be estimated based on available data. Note that in this thesis we only look at the right tail of the distribution, but in some applications the left tail might be of interest.

Initial work on EVT was performed by Maurice Fréchet (1927), Ronald Fisher and Leonard Tippett (1928), Richard von Mises (1936) and Boris Gnedenko (1943). Their research focused on the behaviour of the sample maximum $X_{n,n}$ and a key result is the Fisher-Tippett-Gnedenko theorem: if there exist normalising constants $a_n > 0$ and b_n such that for all x,

$$\lim_{n\to\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = G(x), \tag{1.1}$$

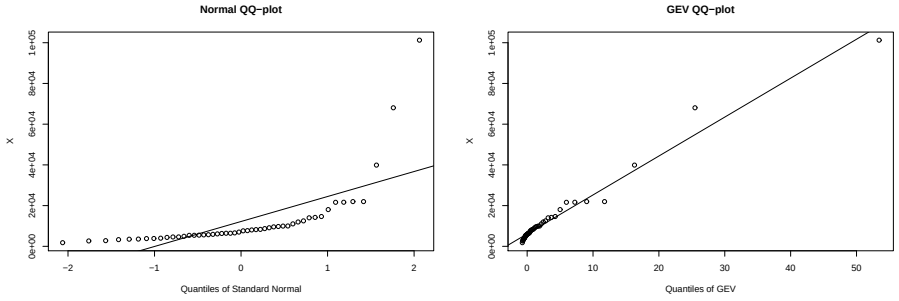
for some non-degenerate distribution function G, then G is necessarily of extreme value type. This means that, up to an affine change of variables, G is the cumulative distribution function (CDF) of the generalised extreme value (GEV) distribution:

$$G(x) = G_\xi(x) = \exp\left(-(1+\xi x)^{-1/\xi}\right) \text{ if } x > -1/\xi.$$

The real parameter ξ is called the extreme value index (EVI). In Section 3.2, we will characterise distributions for which sequences $\{a_n\}$ and $\{b_n\}$ exist such that (1.1) holds.

As an example we look at the 50 most expensive wines of the world. For each of the wines, we obtained the maximum price per bottle over several years from `http://www.wine-searcher.com/most-expensive-wines`. When

making a normal QQ-plot of the maximum prices, we see a convex shape meaning that the normal distribution underestimates the upper tail of the distribution of the maxima. Based on the Fisher-Tippett-Gnedenko theorem, we expect that the maxima follow a GEV distribution. The GEV QQ-plot with $\xi = 1.025$, which was obtained by fitting the GEV to all 50 maxima using maximum likelihood estimation (MLE), shows that the GEV provides a much better fit for the data. When estimating the probability that the maximum wine price exceeds the largest observed maximum wine price, i.e. 101 251, we obtain estimates $1.96 \times 10^{-11}\%$ using the fitted normal distribution and 9.34% using the fitted GEV distribution. This illustrates the large difference between using the normal distribution vs. EVT techniques to provide estimates for quantities regarding the tail of the distribution.



(a) (b)

Figure 1.2: Wine data: (a) normal QQ-plot and (b) GEV QQ-plot of maximum wine prices.

The Fisher-Tippett-Gnedenko theorem shows that EVT is different from classical theory dominated by the central limit theorem (CLT), and that it hence requires separate treatment. An early important reference is Gumbel (1958) which gives the first main overview of extreme value theory and its asymptotic results. However, before 1970 research on extremes was still limited. The fundamental work by Laurens de Haan in his PhD thesis (1970), and by Guus Balkema and Laurens de Haan (1974) and James Pickands III (1975) provided the foundations for new theoretical developments in EVT. Their work focused on the probabilistic and stochastic properties of sample extremes, thus moving away from the sample maximum. More details on (univariate) EVT can be found in Section 3.2. There we also formulate the Pickands–Balkema–de Haan theorem (3.3) which is a basic building block for EVT.

Important applications of extreme value theory can be found in climatology, engineering, geophysics and hydrology, see Chapter 1 in Beirlant et al. (2004)

for an overview. In this thesis, we will focus on applications of EVT that are of interest to finance and insurance.

Stock market crashes such as Black Monday in 1987 and the 2007–2008 financial crisis are examples of extreme events in finance. Bankers and risk managers want to assess and hedge their risks, and price financial instruments taking these risks into account. Therefore, they need models that describe extreme events appropriately. Important quantities for risk assessment are for example return periods of large events, Value-at-Risk (VaR) and Tail Value-at-Risk (TVaR), and excesses over large thresholds.

EVT has been widely used in non-life insurance for years since large insurance claims can be a threat to the solvency of the company. Loss models are needed to set suitable premiums, calculate risk measures and determine capital requirements for solvency regulations. This needs to ensure that the company remains solvent, even in the case of catastrophes. Typical examples of such catastrophes are earthquakes, floods, industrial fires and plane crashes, but also automobile insurance can lead to large claims.

Since insurance companies want to cover themselves against large losses, they often purchase reinsurance. This is a contract where the reinsurer covers part of the insurance risk of the client (which is typically an insurance company). In this way, part of the risk of the insurance company is transferred to the reinsurer. A typical example of a reinsurance contract is excess-loss insurance where the reinsurer covers the client's losses above a certain retention level. Hence, the modelling of extreme events is crucial for reinsurers. For an introduction to reinsurance and its actuarial and statistical aspects we refer to Albrecher et al. (2017). Another example of risk transfer is a catastrophe (CAT) bond which is used by insurance companies to protect themselves against large losses caused by natural disasters such as hurricanes. This bond has higher coupons than a standard bond. However, in case of a catastrophe, the coupons or even the principal are not paid to the investor since they are used to provide extra capital for the insurer to cover the catastrophe losses. This is an example where modelling of extreme events is important from both a financial and actuarial point of view. We refer to Embrechts et al. (1997) for more details on the application of EVT to finance and insurance.

In financial risk management, a Black Swan refers to an event that is deemed improbable yet has massive consequences. In Chapter 3 we propose a way to investigate if the 2007–2008 financial crisis was a Black Swan event for a given bank based on weekly log-returns. More specifically, using techniques from extreme value methodology we compare the tail behaviour of the negative log-returns for two specific horizons:

- Pre-crisis: from 1 January 1994 until 7 August 2007.
- Post-crisis: from 8 August 2007 until 23 September 2014.

We illustrate this approach with Barclays and Credit Suisse data, and argue that Barclays can be considered as having experienced a Black Swan event whereas this is not the case for Credit Suisse. We then link the differences in tail risk behaviour between these banks with economic indicators. We emphasise the use of statistical methods for modelling univariate extremes linked with graphical support.

To predict possible damage by earthquakes, the earthquake engineering community, disaster management agencies and the insurance industry need models for earthquake magnitudes, and especially estimates for the area-characteristic, maximum possible earthquake magnitude T_M. Davies and Kijko (2003) estimate probabilities of certain damages, for different building types, based on probabilistic models for earthquake magnitudes. Kijko et al. (2015) estimate the maximum possible magnitude for the Cape Town area and model the expected damage to buildings for an earthquake of this size, i.e. the worst-case scenario. A widely used parametric model for earthquake magnitudes M is the Gutenberg-Richter (GR) distribution which is a (doubly) truncated exponential distribution with survival function

$$P(M > m) = \frac{e^{-\beta m} - e^{-\beta T_M}}{e^{-\beta t_M} - e^{-\beta T_M}}, \text{ for } t_M < m < T_M. \tag{1.2}$$

In Chapter 5, we consider the specific case of the Dutch province of Groningen where earthquakes are induced by gas extraction since the end of the 1980s. Up to the end of 2016, 286 earthquakes with magnitudes larger than 1.5 occurred in Groningen with a maximum magnitude of 3.6. In Figure 1.3 we make an exponential QQ-plot of the magnitudes of these earthquakes. The fitted line using the lower truncated exponential distribution ((1.2) with $T_M = +\infty$) indicates that an unbounded model does not make sense here as the QQ-plot bends off near the largest observations. The fitted line using the GR distribution with $T_M = 3.83$ clearly models the data better.

This truncation effect is not only seen in earthquake magnitudes. In several applications, ultimately at the largest data, truncation effects can be observed when analysing tail characteristics of statistical distributions. This means that we observe realisations of the random variable X with $X =_d Y \mid Y < T$. Here, Y is the parent variable of X and T is the endpoint of X. In the Gutenberg-Richter distribution, the underlying distribution is the (lower truncated) exponential distribution.

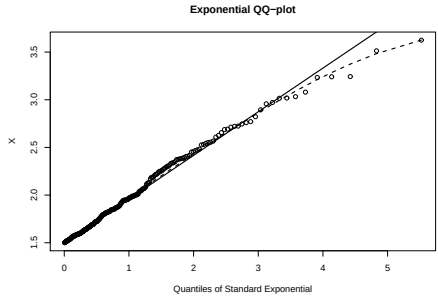


Figure 1.3: Exponential QQ-plot of magnitudes in Groningen with fit based on lower truncated exponential distribution (full line) and GR distribution (dashed line).

Recently, Aban et al. (2006) and Beirlant et al. (2016a) discussed tail fitting for truncated Pareto-type distributions, i.e. the parent variable Y is of Pareto-type. However, as is the case for the Gutenberg-Richter distribution, in some applications the underlying distribution appears to have a lighter tail than the Pareto distribution. In Chapter 4 we generalise the classical peaks over threshold (POT) approach for distributions with EVI $\xi > -1/2$ to allow for truncation effects. This enables a unified treatment of extreme value analysis (EVA) for truncated heavy and light tails. We use a pseudo maximum likelihood approach to estimate the model parameters and consider extreme quantile estimation. The new approach is illustrated on examples from hydrology and geophysics. Moreover, we perform simulations to illustrate the potential of the method on truncated heavy and light tails.

The techniques from Chapter 4 can then be used to estimate the maximum possible earthquake magnitude T_M. In Chapter 5, we also look at two other EVT-based endpoint estimators, and at estimators that are used in the geophysical literature (see e.g. Kijko and Singh, 2011). Next to estimates for the endpoint, we also consider upper confidence bounds to quantify uncertainty of the point estimates. We apply the techniques to provide estimates, and upper confidence bounds, for the maximum possible earthquake magnitude in Groningen. Furthermore, we compare the methods from extreme value theory and the geophysical literature through simulations.

In risk analysis, a global fit that appropriately captures the body and the tail of the distribution of losses is essential. Modelling the whole range of the losses using a standard distribution is usually very hard and often impossible due to the specific characteristics of the body and the tail of the loss distribution. A

possible solution is to combine two distributions in a splicing model (Klugman et al., 2012): a light-tailed distribution for the body which covers light and moderate losses, and a heavy-tailed distribution for the tail to capture large losses.

Financial risk data are often censored and/or truncated, see e.g. Klugman et al. (2012) and Antonio and Plat (2014). A common source of lower truncation in insurance is a deductible. Claims below this threshold are not reported to the insurer since nothing will be paid to the insured if the loss is below the deductible. It can take a long time before an insurance claim is closed and the final cost of the claim is thus not always known at the moment of evaluation. Before the claim is closed, the final claim cost is right censored with the payment up to date as lower bound.

In Chapter 6, we propose a splicing model with a mixed Erlang (ME) distribution for the body and a Pareto distribution for the tail. This combines the flexibility of the ME distribution with the ability of the Pareto distribution to model extreme values. We thus avoid ad hoc combinations of a standard light-tailed distribution, such as the lognormal or the Weibull distribution, for the body with a heavy-tailed distribution for the tail, as proposed in many papers. We extend our splicing approach to censored and/or truncated data where the fitting procedure makes use of the expectation-maximisation (EM) algorithm. Using practical examples from (re)insurance, we illustrate the flexibility of this splicing model.

In this dissertation we do not look at the robustness properties of EVT estimators. However, it is important to note that robust statistics can be used to improve EVA. Dell'Aquila and Embrechts (2006) show that robust statistics can be used to identify influential observations, and detect deviating substructures or model misspecification. In the literature, several robust versions of EVT estimators have been proposed, see e.g. Dupuis and Field (1998), Vandewalle et al. (2007) and Hubert et al. (2013).

Part I

Robust Sparse PCA

Chapter 2

Sparse PCA for high-dimensional data with outliers

This chapter is based on

Hubert, M., Reynkens, T., Schmitt, E. and Verdonck, T. (2016). Sparse PCA for High-Dimensional Data With Outliers. *Technometrics*, **58**(4), 424–434.

2.1 Introduction

Principal component analysis (PCA) is a popular technique used for dimension reduction. The idea is to find a number of uncorrelated linear combinations of the original variables that capture most of the covariance structure of the original data. These combinations are called the principal components (PCs). Those directions are chosen such that they are orthogonal and sequentially maximise the variance of the projected data. Typically one does not use all the PCs, but only the first k explaining a sufficient portion of the total variance (i.e. information) of the original data. Despite its advantages, classical principal component analysis (CPCA) also has several drawbacks; two of which we will focus on.

First, CPCA often results in PCs that are difficult to interpret because most of the loadings are neither very small nor very large in absolute value. To increase

interpretability, sparse PCA methods were developed to estimate PCs with many zero loadings. This is useful when the data is high-dimensional, since only a subset of the original variables may need to be analysed or measured. Two popular methods for performing sparse PCA are SCoTLASS (Jolliffe et al., 2003) and SPCA (Zou et al., 2006).

Second, it is well known that outliers present in the data can heavily affect the CPCA estimates. Several robust alternatives for CPCA have been proposed including a projection pursuit principal component analysis (PP-PCA) approach (Li and Chen, 1985; Hubert et al., 2002; Croux and Ruiz-Gazen, 2005), spherical PCA (Locantore et al., 1999), and ROBPCA (Hubert et al., 2005).

We propose a new method, RObust Sparse Principal Component Analysis (ROSCPA), combining the advantageous properties of sparse and robust PCA. Previous work on this problem has been done by Croux et al. (2013), who developed a sparse version of the robust PP-PCA method by integrating sparsity principles into the formulation of PP-PCA. Since we believe that the detection of outliers may be the more difficult, and crucial, challenge, we approach the problem from a different direction, and develop a sparse modification of the robust ROBPCA method. The main difference is that we partially separate the outlier detection step from the sparsification step. As we detail later, doing so results in greater robustness and more accurate sparse estimates.

Note that our model assumptions are different from those studied in Candès et al. (2011) and Zhou et al. (2010). Whereas we are searching for a subspace spanned by sparse vectors, in the latter papers not the subspace but the errors are supposed to be sparse. This allows to recover the subspace exactly with a convex optimisation program.

In Section 2.2 we first give a summary of existing methods for sparse and/or robust PCA, and then we detail our new method together with a new criterion to select the sparsity parameter. Section 2.3 contains the results of a simulation study, whereas Section 2.4 illustrates ROSPCA on a real dataset. In Section 2.5, we discuss an extension of ROSPCA for skewed data, and we apply it to a financial data example. Finally, Section 2.6 contains conclusions and directions for further research.

2.2 Methods

2.2.1 Classical PCA

To fix notation, we begin by defining PCA for a data matrix, $\boldsymbol{X} = \boldsymbol{X}_{n,p} \in \mathbb{R}^{n \times p}$. In general, the subscripts denote the dimensions of the matrix and will only be added when appropriate. The p-dimensional observations in $\boldsymbol{X}$ are denoted by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. The loadings of the PCs, i.e. the components of the linear combinations, are in the columns of the orthogonal *loadings matrix* $\boldsymbol{P}$. Given estimated loadings $\boldsymbol{P}$ and centre $\hat{\boldsymbol{\mu}}$, projecting the centred $\boldsymbol{X}$ on the new directions yields the *scores matrix* $\boldsymbol{T} = (\boldsymbol{X} - \mathbf{1}_n \hat{\boldsymbol{\mu}}')\boldsymbol{P}$, with $\mathbf{1}_n$ a column vector consisting of n ones.

Classical PCA can be described as searching for a $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{P}$ such that the scores have maximal variance, and are uncorrelated. The PCA directions then correspond to the eigenvectors of the classical covariance matrix $\boldsymbol{S}$ of $\boldsymbol{X}$, whereas the variance of the data projected on an eigenvector is equal to the corresponding eigenvalue of $\boldsymbol{S}$. Note that when the variances of the original variables differ greatly, the data should first be standardised. If one uses the componentwise standard deviation, this comes down to computing the eigenvectors of the correlation matrix of $\boldsymbol{X}$.

Typically, $k \ll p$ dimensions are needed to express the information in the data. Various approaches exist to select the number of components to retain, k. One of the simplest and most popular is the *scree plot*. It plots the sorted, decreasing eigenvalues versus their index. The number of components corresponding to the point at which an elbow in the plot occurs is then selected. Following the selection of the number of components, only the first k columns of $\boldsymbol{P}$ are used and denoted as $\boldsymbol{P}_{p,k} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k]$.

2.2.2 Sparse PCA

Sparse PCA has the advantage of making the interpretation of the PCs easier. A simple way to accomplish this is to set all loadings with absolute value smaller than a certain threshold to zero. This method is called *simple thresholding*. Cadima and Jolliffe (1995) noticed that this method can be potentially misleading. For example, one should also look at the standard deviations of variables to determine the contribution of a variable to a certain PC.

To overcome the issues of that early method, a number of methods have been developed. One of these is simplified component technique - LASSO

(SCoTLASS), which was proposed by Jolliffe et al. (2003). It integrates an L_1 constraint with PCA, yielding sparse loadings. The resulting objective function seeks the orthogonal loadings $\boldsymbol{p}_j$ maximising the variance explained by the fitted model, subject to the constraint $\|\boldsymbol{p}_j\|_1 \leqslant \eta_j$, a sparsity constraint, where $\|\boldsymbol{p}_j\|_1$ is the L_1 norm of $\boldsymbol{p}_j$. We will work with the dual of this problem:

$$\boldsymbol{p}_j = \underset{\|\boldsymbol{p}\|=1, \boldsymbol{p} \perp \boldsymbol{p}_1, \ldots, \boldsymbol{p} \perp \boldsymbol{p}_{j-1}}{\operatorname{argmax}} \boldsymbol{p}' \boldsymbol{S} \boldsymbol{p} - \lambda_j \|\boldsymbol{p}\|_1, \tag{2.1}$$

where $\boldsymbol{p}_j$ is the jth PCA direction. Under this formulation, λ_j is the sparsity parameter for SCoTLASS, in place of η_j. A higher value of λ_j corresponds to greater sparsity, and a value of zero corresponds to no sparsity.

2.2.3 Robust PCA

The loadings matrix estimated by CPCA and sparse PCA is very sensitive to outliers. Robust principal component analysis addresses this issue. Two well known robust PCA methods are robust Projection Pursuit PCA (PP-PCA) and ROBPCA. PP-PCA maximises a robust measure of spread to obtain consecutive directions on which the data is projected. Croux and Ruiz-Gazen (2005) proposed a version that serves as the basis for one variant of sparse, robust PCA. The ROBPCA method (Hubert et al., 2005) combines ideas from projection pursuit and robust covariance estimation. These approaches will be discussed in greater detail below, when we encounter sparse versions.

To detect PCA outliers, two notions of distance are used: robust score distances and orthogonal distances. The *robust score distance (SD)* measures the robust statistical distance from a PC score to the centre of the scores. For an observation $\boldsymbol{x}_i$, the robust score distance is defined as

$$\mathrm{SD}_i = \sqrt{\sum_{j=1}^{k} \frac{(\boldsymbol{t}_i)_j^2}{l_j}} = \sqrt{\boldsymbol{t}_i' \boldsymbol{L}^{-1} \boldsymbol{t}_i}, \tag{2.2}$$

with k the number of PCs, $(\boldsymbol{t}_i)_j$ the jth component of the ith score $\boldsymbol{t}_i$ and $\boldsymbol{L}$ the diagonal matrix containing the robust eigenvalues corresponding to the robust PCs. We set the cut-off for observations with high SD values at $c_{\mathrm{SD}} = \sqrt{\chi^2_{k,0.975}}$, the square root of the 97.5% quantile of a chi-squared distribution with k degrees of freedom. This is justified when the scores are approximately normally distributed.

The *orthogonal distance (OD)* of an observation $\boldsymbol{x}_i$ to the PCA subspace is given by

$$\mathrm{OD}_i = \|\boldsymbol{x}_i - \hat{\boldsymbol{\mu}} - \boldsymbol{P}_{p,k} \boldsymbol{t}_i\|. \tag{2.3}$$

Note that $\hat{\boldsymbol{\mu}} + \boldsymbol{P}_{p,k}\boldsymbol{t}_i$ is the projection of $\boldsymbol{x}_i$ on the PCA subspace determined by $\boldsymbol{P}_{p,k}$ and $\hat{\boldsymbol{\mu}}$. To obtain a cut-off for the orthogonal distances, we follow the approach taken in Hubert et al. (2005). This makes use of the Wilson-Hilferty approximation for a chi-squared distribution, which implies that the orthogonal distances to the power 2/3 are approximately normally distributed. To obtain estimates of the centre and scale of this distribution we use the univariate minimum covariance determinant (MCD) (Rousseeuw, 1984), a robust estimator that searches for the subset of size $\frac{n}{2} < h \leqslant n$ that has the smallest variance and bases location ($\hat{\mu}_{MCD}$) and scale ($\hat{\sigma}_{MCD}$) estimates on it. Given these parameters, the cut-off is defined as $c_{\text{OD}} = (\hat{\mu}_{\text{MCD}} + \hat{\sigma}_{\text{MCD}} z_{0.975})^{3/2}$, with $z_{0.975}$ the 97.5% quantile of the standard normal distribution.

2.2.4 SRPCA

Croux et al. (2013) proposed a robust, sparse method that combines ideas from the PP-PCA approach and sparse PCA. It will be used as a benchmark in our simulations and a real data example. Their approach consists of adding the L_1 penalty into the PP-PCA equations. The method thus looks for directions that maximise the scale of the data projected on them under the constraint that the loadings of these directions should not be too large. The jth sparse PCA direction is given by

$$\tilde{\boldsymbol{p}}_j = \begin{cases} \underset{\|\boldsymbol{p}\|=1}{\operatorname{argmax}} \; S(\boldsymbol{p}'\boldsymbol{x}_1, \ldots, \boldsymbol{p}'\boldsymbol{x}_n) - \lambda_1 \|\boldsymbol{p}\|_1 & \text{if } j = 1 \\ \underset{\|\boldsymbol{p}\|=1,\, \boldsymbol{p} \perp \tilde{\boldsymbol{p}}_1, \ldots, \boldsymbol{p} \perp \tilde{\boldsymbol{p}}_{j-1}}{\operatorname{argmax}} \; S(\boldsymbol{p}'\boldsymbol{x}_1, \ldots, \boldsymbol{p}'\boldsymbol{x}_n) - \lambda_j \|\boldsymbol{p}\|_1 & \text{if } 1 < j \leqslant p, \end{cases} \tag{2.4}$$

where S is a measure of scale. If one uses the sample standard deviation for S, this method is nothing more than SCoTLASS. To obtain robust principal components, Croux et al. (2013) suggest to use the robust Q_n estimator of scale (Rousseeuw and Croux, 1993). The Q_n is the first quartile of the pairwise distances between the elements of a vector. The data are typically centred using a robust estimator for the centre (e.g. using the L_1-median). Then, one applies the PP-PCA steps on the $\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}$ (for $1 \leqslant i \leqslant n$), with $\hat{\boldsymbol{\mu}}$ the robust estimate for the centre.

The sparsity parameter λ_j can vary across the different PCs. Croux et al. (2013) make the relative importance of the L_1 penalty comparable across the different PCs. This means that there is a similar degree of sparsity across the PCs. They take $\lambda_j = \lambda v_j$ where v_j can be defined as follows. Suppose we have found the $j-1$ first PC directions and denote by $\boldsymbol{X}_j^{\perp}$ the data projected on the space orthogonal to the space spanned by the $j-1$ first PC directions. The number v_j is then the average of the variance measure S^2 applied to the columns of $\boldsymbol{X}_j^{\perp}$.

Note that v_1 is the average of the variance measure S^2 applied to the columns of $\boldsymbol{X}$. This definition is used in the R packages *pcaPP* (Filzmoser et al., 2014) and *rrcovHD* (Todorov, 2014) and differs slightly from the definition in Croux et al. (2013). Hence, there is only one tuning parameter to select: the sparsity parameter λ. We denote this method by SRPCA as in Todorov and Filzmoser (2013).

To find the sparse PCA directions in (2.4), the expressions need to be maximised over a p-dimensional space. This optimisation problem is non-convex. The Grid algorithm of Croux et al. (2007) is an accurate algorithm that is used to obtain the PCA directions in the PP-PCA approach. In Croux et al. (2013), the authors extend it for sparse PCA and provide a detailed description of the algorithm. Since SRPCA is a generalisation of the PP-PCA approach, Croux et al. (2013) proposed to extend the Grid algorithm to compute the sparse directions. Henceforth, we will use this algorithm to compute the sparse loadings of SCoTLASS and SRPCA. By default, the maximum number of iterations is equal to 10, but we noticed that the algorithm does not yet converge then. We use a maximum of 75 iterations instead which provides stable results.

2.2.5 ROSPCA

Hubert et al. (2005) proposed a robust PCA algorithm combining ideas from projection pursuit and the MCD estimator, which they called ROBPCA. Many steps in ROBPCA anticipate those of ROSPCA as the robustness properties of the latter derive almost directly from the former. Intuitively, they can be compared as follows. ROBPCA finds an outlier-free subset which determines a robust subspace. Then, it projects the data onto this subspace to estimate the eigenvectors and eigenvalues robustly. The ROSPCA method (RObust Sparse PCA) integrates sparse PCA into ROBPCA. In doing so, ROSPCA finds a subset that determines a robust, *sparse* subspace, and then estimates the eigenvectors and eigenvalues while preserving sparsity.

Not surprisingly the method contains two hyperparameters: α which determines the degree of robustness and λ which regulates the sparsity. The value of α must satisfy $0.5 \leqslant \alpha < 1$ and needs to be chosen in advance. It constitutes a lower bound on the number of regular observations, so at most $100(1-\alpha)\%$ of the n data points are allowed to be outlying. If no a priori information about the amount of outliers is available, we recommend to set $\alpha = 0.5$, yielding maximal robustness. The choice of the sparsity parameter λ will be discussed in Section 2.2.6.

The ROSPCA algorithm consists of an outlier detection part (step 1), and a sparsification part (steps 2 and 3):

1. The first part is similar to ROBPCA, so we describe it only shortly. When a standardisation is appropriate, the variables are first robustly standardised by means of the componentwise median and the Q_n. Then, using the singular value decomposition (SVD) of the resulting data matrix, the p-dimensional data space is reduced to the affine subspace spanned by the n observations. We denote the resulting data matrix (of rank at most $n-1$) by $\tilde{\boldsymbol{X}}$. Next, for each $\tilde{\boldsymbol{x}}_i$ the Stahel-Donoho outlyingness is computed as

$$\operatorname{outl}(\tilde{\boldsymbol{x}}_i) = \max_{\boldsymbol{v} \in B} \frac{|\tilde{\boldsymbol{x}}_i' \boldsymbol{v} - \hat{\mu}_{\mathrm{MCD}}(\tilde{\boldsymbol{x}}_j' \boldsymbol{v})|}{\hat{\sigma}_{\mathrm{MCD}}(\tilde{\boldsymbol{x}}_j' \boldsymbol{v})} \tag{2.5}$$

   where $\hat{\mu}_{\mathrm{MCD}}$ and $\hat{\sigma}_{\mathrm{MCD}}$ are the univariate MCD estimators of location and scale. The set $\boldsymbol{B}$ consists of all directions $\boldsymbol{v}$ passing through two data points (or a random subset of these directions if n is very large).

   Thereafter, the $h_0 = \lceil \alpha n \rceil + 1$ observations with smallest outlyingness are considered, they are mean-centred and SVD is applied to them to find the k-dimensional subspace most closely to them (in L_2-norm). Here, the scree plot can be used to find an appropriate value for k, or the cumulative percent variation (CPV). For example, one could select k such that $\mathrm{CPV} = \sum_{j=1}^{k} s_j^2 / \sum_{j=1}^{p} s_j^2 \geqslant 80\%$ with s_j the singular values of the SVD decomposition. Next, following Engelen et al. (2005), given the orthogonal distances to the preliminary subspace, we consider all observations with ODs smaller than the corresponding cut-off (as explained in Section 2.2.3). This yields an outlier-free index set H_1 of size h_1, which typically will be larger than h_0, in particular when α is chosen much smaller than the real proportion of regular observations.

2. Whereas ROBPCA applies CPCA on the observations from H_1, ROSPCA now uses sparse PCA. More precisely, we first standardise the data points of $\boldsymbol{X}$ with indices in H_1 using the componentwise median and the Q_n. Performing sparse PCA on them, by means of the Grid-based implementation of SCoTLASS with sparsity parameter λ, yields the sparse loadings matrix $\boldsymbol{P}_1 \in \mathbb{R}^{p \times k}$.

   We then perform an additional reweighting step that incorporates information about the sparse structure of the data, forming a bridge between the sparse and robust components of the algorithm and increasing efficiency. We discard variables with zero loadings on all k PCs and we then compute the orthogonal distances to the estimated sparse PCA subspace. This yields an index set H_2 of observations with orthogonal distance smaller than the cut-off corresponding to these new orthogonal distances. We now standardise the subset of $\boldsymbol{X}$ with indices in H_2 using the componentwise median and the Q_n of the observations in H_1 (we use the same standardisation as in

the first time sparse PCA is applied). Then, sparse PCA is applied onto them, again by means of the Grid-based implementation of SCoTLASS with sparsity parameter λ. To get a full loadings matrix $\boldsymbol{P}_2$, we also need to add zero rows for all discarded variables to the estimated loadings matrix. The k-dimensional scores after reweighting are then given by $\boldsymbol{T} = (\boldsymbol{X} - \mathbf{1}_n \hat{\boldsymbol{\mu}}_1')\boldsymbol{P}_2$, with $\hat{\boldsymbol{\mu}}_1'$ the median of the observations in H_1. Intuitively, the goal of this reweighting is to recapture information from observations that are only outlying due to their behaviour on variables that are found to be unimportant in our model, and use this information to obtain better estimates of the loadings corresponding to the important variables. Such observations will still have high OD values since the variables on which they are outlying will be compared to zero loadings in $\boldsymbol{P}_2$.

3. Finally, the eigenvalues are estimated robustly by applying the Q_n^2 estimator on the scores of the observations with indices in H_2. We need to use a robust measure of scale because observations with low OD and high SD that are included can influence the eigenvalue estimation. In order to robustly estimate the centre, we compute the score distances and look at all observations of H_2 with a score distance smaller than the corresponding cutoff, this is the set H_3. We then estimate the centre by the mean of these observations which gives the final centre $\hat{\boldsymbol{\mu}}$ and the final scores $\boldsymbol{T} = (\boldsymbol{X} - \mathbf{1}_n \hat{\boldsymbol{\mu}}')\boldsymbol{P}_2$. We finally recompute the estimates of the eigenvalues by computing the sample variance of the (new) scores of the observations with indices in H_3 (the observations with low OD and high SD are not included anymore). The eigenvalues are sorted in descending order, so the order of the PCs may change. The columns of the loadings and scores matrices are changed accordingly.

Note that when it is not necessary to standardise the data, we only centre the data as in the scheme above, but do not scale them.

2.2.6 Selection of sparsity parameters

SRPCA, SCoTLASS and ROSPCA use a scalar sparsity parameter λ in the Grid algorithm. Croux et al. (2013) select λ using a Bayesian information criterion (BIC) type criterion. It looks at the ratio of residual variances and the degree of sparsity of the loadings matrix. These residual variances are computed by applying the Q_n^2 estimator to the sums of the squared OD statistics of the sparse and unconstrained PCA models. However, in our simulations and real data examples, this BIC approach selects λ values that are noticeably too sparse for ROSPCA, so we only use it for SRPCA. We choose λ by minimising a

BIC-type criterion based on the conventional formulation derived to use the residual sum of squares (RSS). Our BIC-type criterion is:

$$\mathrm{BIC}(\lambda) = \ln\left(\frac{1}{h_1 p}\sum_{i=1}^{h_1}\mathrm{OD}^2_{(i)}(\lambda)\right) + \mathrm{df}(\lambda)\frac{\ln(h_1 p)}{h_1 p}, \tag{2.6}$$

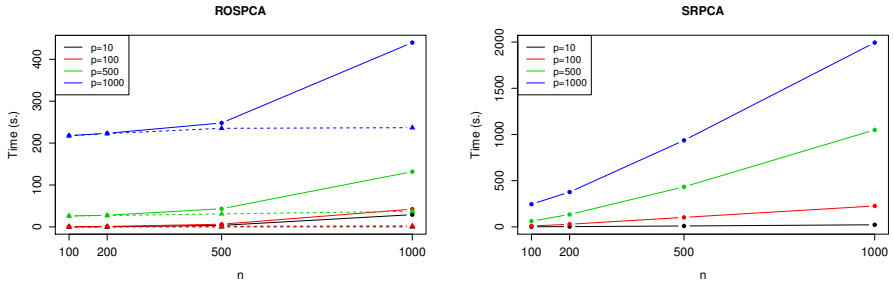
where h_1 is the size of H_1, and $\mathrm{OD}_{(i)}(\lambda)$ is the ith smallest orthogonal distance for the model when using λ as the sparsity parameter. This criterion is similar to the BIC in regression, with the PCA orthogonal distances in place of the regression residuals. In ordinary regression, the residuals are univariate. Because the ODs are norms of p-dimensional vectors, we have to include p in (2.6). Moreover we use h_1 instead of n as this denotes the size of an outlier-free subset which does not depend on λ. After reweighting, if contamination is not high, h_1 is often close to n. Similar to Croux et al. (2013), $\mathrm{df}(\lambda)$ is taken as the number of non-zero loadings when λ is used as the sparsity parameter.

The first part of the criterion measures the quality of the fit whereas the second term penalises for model complexity, reflecting a trade-off between accuracy and sparsity. In practice, we select λ by minimising the BIC over the interval $[0, \lambda_{\max}]$ where $\lambda_{\max}$ gives full sparseness (exactly one non-zero loading per PC). We do this by looking at a grid of (usually equidistant) λ values over this interval.

Note that the computation of the index set H_1 in ROSPCA (step 1) does not depend on the choice of the sparsity parameter. It is therefore not necessary to run the full method each time we compute the BIC for a certain λ value. We perform the parts that are independent of λ only once and we then use this, for each value of λ we look at, as input for the parts that depend on the sparsity parameter (steps 2 and 3). This approach reduces the computation time and can lead to a considerable speed-up if many λ values need to be evaluated. This computational improvement cannot be applied to the SRPCA and SCoTLASS methods because in that case the Grid algorithm fully depends on the value of λ.

The computation time of ROSPCA is the result of its initial outlier detection part (step 1) and the remaining steps 2 and 3 to obtain sparsity. Figure 2.1 displays the computation times in seconds of ROSPCA (left) and SRPCA (right) for a range of values of n and p, and for $k = 2$ and $\lambda = 0$ using R 3.1.1 (`R` Core Team, 2014) on Windows 7 (64-bit) OS with an Intel Core i7-3770 CPU @ 3.40GHz. The ROSPCA plot contains a further breakdown of computation time between the sparse and total computation times. The difference is the computation time attributable to the outlier detection step, which becomes more time consuming as n increases. Both ROSPCA and SRPCA show an increase in computation time as a function of n and p. The effect is noticeably

stronger though for SRPCA, which shows much higher computation times as a function of both parameters (note the difference in the y-axis). This is primarily due to the way that the methods achieve robustness. ROSPCA performs a single outlier detection step, and then in the following steps it calculates the computationally inexpensive standard deviation for each direction in the Grid algorithm. In contrast, SRPCA relies on the comparatively slower Q_n statistic because robustness is achieved at the same time as sparsity is imposed. Note that the computation time of SRPCA is independent of the sparsity parameter λ. For ROSPCA, the computation time will decrease with λ since for higher values of λ, more variables can be excluded in the additional reweighting step which decreases the computation time of the second execution of SCoTLASS. We used $\lambda = 0$ to construct Figure 2.1, so computation times are lower when more sparsity is imposed using a higher value of λ.



(a) (b)

Figure 2.1: Computational performance of (a) ROSPCA and (b) SRPCA for varying values of n and p. The ROSPCA plot displays both the sparse (dashed line) and total (solid line) computation times.

2.3 Simulations

2.3.1 Layout of the simulation study

To evaluate the robustness, accuracy and sparsity of ROSPCA, we compare its performance with that of SRPCA, SCoTLASS, CPCA and ROBPCA on outlier-free and contaminated data. In specifying our simulations, we generate data from a multivariate normal distribution with a covariance matrix that has sparse eigenvectors. A varying proportion of the observations are replaced with outliers in order to test the robustness of the methods. We first standardise

the data so that performing CPCA results in computing the eigenvectors (and -values) of the correlation matrix. Therefore, we need to generate a correlation matrix with sparse eigenvectors. First, we give a detailed description of the setup. Next, we evaluate the accuracy of the different PCA methods on the simulated data using performance measures based on the estimated loadings.

Let $\mathbb{R}^p$, with $p \geqslant 8$, be our original data space, and let $k = 2$ be the number of important components. We generate a correlation matrix such that it has sparse eigenvectors. We design the correlation matrix to have 3 groups of variables with no correlation between variables from different groups. The first two groups consist of b variables each, where b is an integer that we choose to be at least 4. The correlation between the different variables of the group is equal to $a_1 \in [-1, 1]$ for group 1 and $a_2 \in [-1, 1]$ for group 2. The third group contains the remaining $p - 2b$ variables, which we specify to be uncorrelated. Our correlation matrix $\boldsymbol{R}$ is thus equal to

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}(a_1) & \mathbf{0}_{b\times b} & \mathbf{0}_{b\times(p-2b)} \\ \mathbf{0}_{b\times b} & \boldsymbol{R}(a_2) & \mathbf{0}_{b\times(p-2b)} \\ \mathbf{0}_{(p-2b)\times b} & \mathbf{0}_{(p-2b)\times b} & \boldsymbol{I}_{p-2b} \end{pmatrix}$$

with $\boldsymbol{R}(x)$ the $b \times b$-matrix with ones on the diagonal and off-diagonal elements $x \in [-1, 1]$, and $\boldsymbol{I}_{p-2b}$ the $(p - 2b)$-dimensional identity matrix. When $a_1 > a_2$, the first two sparse eigenvectors are given by $\boldsymbol{p}_1 = -\frac{1}{\sqrt{b}}\boldsymbol{q}_1$ and $\boldsymbol{p}_2 = -\frac{1}{\sqrt{b}}\boldsymbol{q}_2$ with $\boldsymbol{q}_1 \in \mathbb{R}^p$ a vector with the first b elements equal to one and zero elsewhere, and $\boldsymbol{q}_2 \in \mathbb{R}^p$ a vector with the second b elements equal to one and zero elsewhere. The first b variables should therefore have zero loadings for the second PC, and similarly for the next b variables and the first PC. It is also clear that the variables from the last group should have zero loadings for both PCs. The order of the first two eigenvectors is changed when a_1 is smaller than a_2. The statements about the zero loadings can be adapted accordingly. Note that the eigenvectors are, neglecting their order, independent of the choice of a_1 and a_2.

Next, the correlation matrix $\boldsymbol{R}$ is transformed into the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{V}^{\frac{1}{2}}\boldsymbol{R}\boldsymbol{V}^{\frac{1}{2}}$, where $\boldsymbol{V}$ is the diagonal matrix containing the variances of the variables to be detailed later. The n observations are generated from a p-variate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. Standard normally distributed noise terms are also added to each of the p variables to make the sparse structure of the data harder to detect. This gives a dataset $\boldsymbol{X} = \boldsymbol{X}_u + \boldsymbol{X}_{noise}$ with $\boldsymbol{X}_u \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{X}_{noise} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{I}_p)$. Finally, $100\varepsilon\%$ of the data points are randomly replaced by outliers. We consider different proportions of outliers, namely $\varepsilon = 0, 0.1, 0.2, 0.3, 0.4$. These outliers are generated from a p-variate normal distribution $\mathcal{N}_p(\boldsymbol{\mu}_{out}, \sigma^2_{out}\boldsymbol{I}_p)$ with $\boldsymbol{\mu}_{out} = 25(0, -4, 4, 2, 0, 4, -4, 2, 3, -3, \ldots, 3, -3)'$ and $\sigma^2_{out} = 20$, as in Croux et al. (2013). Importantly, these outliers do not follow the correlation structure

determined by $\boldsymbol{R}$. They will therefore bias non-robust sparse methods trying to estimate the sparse structure. We also denote the dataset with the outliers by $\boldsymbol{X}$.

First, we consider a low-dimensional setting with $p = 10$ dimensions and $b = 4$ in our simulations, so we have two blocks of four useful variables and the last two variables are noise. We take $a_1 = 0.9$ and $a_2 = 0.5 < a_1$ which gives eigenvalues $3.7, 2.5, 1, 1, 0.5, 0.5, 0.5, 0.1, 0.1, 0.1$ and the first two eigenvectors of $\boldsymbol{R}$ are given by $\boldsymbol{p}_1 = -\frac{1}{2}(1, 1, 1, 1, 0, 0, 0, 0, 0, 0)'$ and $\boldsymbol{p}_2 = -\frac{1}{2}(0, 0, 0, 0, 1, 1, 1, 1, 0, 0)'$. Importantly, the difference between the first and second eigenvalue is large enough such that the methods can clearly determine that $\boldsymbol{p}_1$ is the loading vector of the first PC. When taking a_1 and a_2 closer together, the difference between the first two eigenvalues gets smaller, so it becomes more difficult for the PCA method to identify which of the first two eigenvectors corresponds to the first PC. We also need to make sure that a_2 is large enough, otherwise the difference between the second and third eigenvalue is too small. This can again cause problems because the PCA method can sometimes select the third eigenvector as the loading vector corresponding to the second PC, making our bias criterion become difficult to interpret. With our choices for a_1 and a_2, the difference between the eigenvalues is large enough to avoid these problems. We take $\boldsymbol{V} = \text{diag}(100, \ldots, 100, 25, \ldots, 25, 4, 4)$, so the variables in a group have the same variance. For each simulated scenario, we generate 500 datasets following the above scheme to thoroughly characterise the behaviour of the methods.

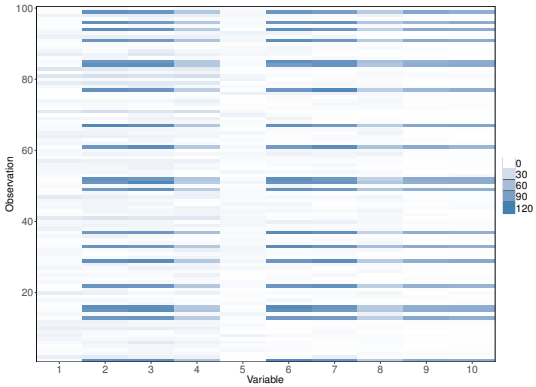


Figure 2.2: Heat map of absolute value of simulated data with $p = 10$, $n = 100$ and $\varepsilon = 0.2$. Outliers are visible in dark blue.

Figure 2.2 shows a heat map of the absolute values of one dataset from our simulation setting with $p = 10$, $n = 100$ and $\varepsilon = 0.2$. The outliers are visible as the observations with values taking a dark blue colour. Despite being fairly easy

to identify on a heat map, we shall see that these can pose difficulties for sparse PCA methods that are not highly robust. We note that the configurations we use to evaluate the considered methods are known to be particularly challenging for them, while they are capable of easily identifying outliers in other configurations that are not clearly revealed by a heat map.

We also look at a high-dimensional setting with $p = 500$ and $k = 2$. In contrast to the low-dimensional setting, the first two groups consist of $b = 20$ variables each, which results in 40 useful variables and 460 noise variables. In the new setting, the eigenvalues are $18.1, 10.5$, 1 (460 times), 0.5 (19 times) and 0.1 (19 times), where we take $a_1 = 0.9$ and $a_2 = 0.5 < a_1$ again. The first two sparse eigenvectors are given by $\boldsymbol{p}_1 = -\frac{1}{\sqrt{20}}\boldsymbol{q}_1$ and $\boldsymbol{p}_2 = -\frac{1}{\sqrt{20}}\boldsymbol{q}_2$ with $\boldsymbol{q}_1 \in \mathbb{R}^{500}$ a vector with the first 20 elements equal to one and zero elsewhere, and $\boldsymbol{q}_2 \in \mathbb{R}^{500}$ a vector with the second 20 elements equal to one and zero elsewhere. We use the same variances for the groups as before: 100 for group 1, 25 for group 2 and 4 for group 3. For each scenario, we now generate 100 datasets following the high-dimensional scheme to keep computations reasonable.

To compare the robustness of the methods, we look at the second principal angle between the subspace spanned by the two dominant eigenvectors of the correlation matrix $\boldsymbol{R}$ and the subspace spanned by the columns of the estimated loadings matrix (the PCA subspace), as was also done in Hubert et al. (2005) and Todorov and Filzmoser (2013). We compute this angle using the algorithm of Björck and Golub (1973). This angle lies between 0 and $\frac{\pi}{2}$, and we divide it by $\frac{\pi}{2}$ to get values between 0 and 1. In the remainder we will refer to the standardised version as the "angle". It is clear that we want values close to 0.

All simulations were performed in R 3.1.1 using following functions: `prcomp` (CPCA), `PcaHubert` (ROBPCA) from the *rrcov* package (Todorov and Filzmoser, 2009) and `SPcaGrid` (SRPCA and SCoTLASS) from *rrcovHD* (Todorov, 2014). We used a self-written function for ROSPCA, based on the code for `PcaHubert`, which is included in the *rospca* package (Reynkens, 2017). For ROSPCA and ROBPCA the parameter α is set to 0.5, yielding maximal robustness. First, we compare the estimation of the PCA subspace and the degree of sparsity attained. Then, we discuss the behaviour of the λ selection step of these algorithms following our BIC-type criterion (2.6) for ROSPCA and SCoTLASS, and the BIC criterion of Croux et al. (2013) for SRPCA.

2.3.2 Results of the simulation study

Subspace estimation

We start with the low-dimensional simulations ($p = 10$). For each simulation setting and each sparse method we report two results as boxplots. On the left is a boxplot of the angle values corresponding to a model fitted by a method with λ selected using the previously discussed criteria. We consider following grid of λ values: $\{0, 0.02, \ldots, 2.5\}$. The boxplot on the right is based on the minimal angle value attained by each method over the same range of λ values. These results provide two insights. First, the boxplot based on the minimal angle values gives a sense of the performance of each method if λ were selected to give the fit closest to the real structure of the data possible for that method. Secondly, this boxplot and the boxplot to its left, based on results from models using λ values selected by a criterion, together give a sense of how successful the information criterion is in selecting an optimal value of λ for the method. For CPCA and ROBPCA, we only have the boxplot of the angle values corresponding to the fitted model.

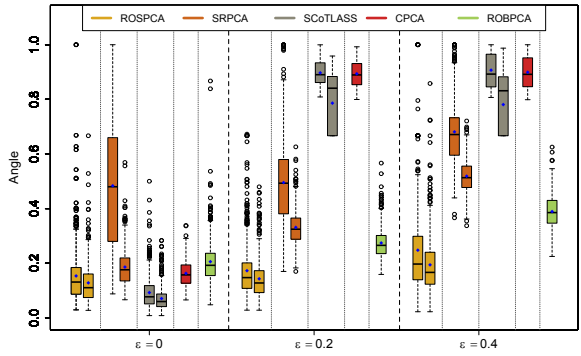


Figure 2.3: Angle values of ROSPCA, SRPCA, SCoTLASS, CPCA and ROBPCA at $p = 10$ and $\varepsilon = \{0, 0.2, 0.4\}$ for $n = 50$.

Figures 2.3, 2.4 and 2.5 show boxplots on datasets of increasing size n and contamination rate ε. Mean values are indicated with blue diamonds. As expected, bias decreases and the angles become less dispersed when n increases. SCoTLASS reports the best results for $\varepsilon = 0$ but performs very badly when contamination is present. Also of note, the boxplots corresponding to models based on selected λ values are only slightly higher than the boxplots based on

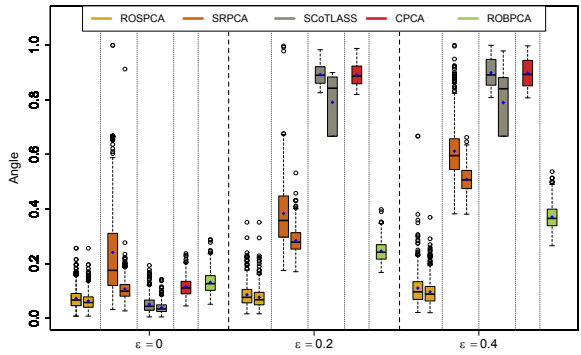


Figure 2.4: Angle values of ROSPCA, SRPCA, SCoTLASS, CPCA and ROBPCA at $p = 10$ and $\varepsilon = \{0, 0.2, 0.4\}$ for $n = 100$.

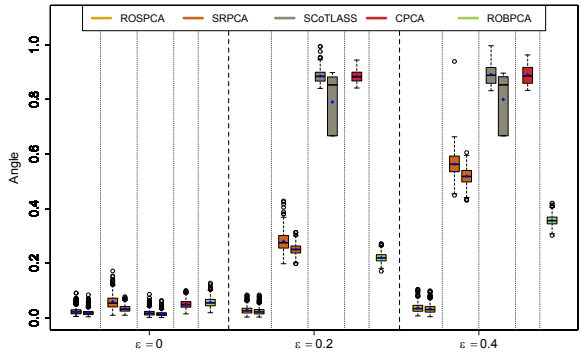


Figure 2.5: Angle values of ROSPCA, SRPCA, SCoTLASS, CPCA and ROBPCA at $p = 10$ and $\varepsilon = \{0, 0.2, 0.4\}$ for $n = 500$.

the minimal angle values, showing that the λ selection problem is tractable for SCoTLASS under these settings. Over all contamination levels, ROSPCA shows a low mean and median bias, even for the case where $\varepsilon = 0.4$. Like SCoTLASS when it is applied to uncontaminated data, the boxplots based on selected λ values and the minimal angles tend to be close, meaning that for ROSPCA, λ is typically selected accurately. At small sample sizes, quite some variability is still present in the estimates, but this decreases substantially at larger sample sizes. In contrast to ROSPCA, SRPCA returns distinctly higher biases, even for

the best possible λ value. Its bias at outlier-free data only becomes reasonably small when n is very large. Furthermore, the difference in the boxplot pairs for SRPCA reveals that the BIC selection criterion proposed by Croux et al. (2013) yields angles that are on average quite distinct from the optimal ones that could be obtained. CPCA is outperformed by the sparse methods SCoTLASS and ROSPCA at outlier-free data, and completely breaks down at contaminated ones. ROBPCA shows an increased bias when contamination is present. A closer look at the results revealed that the method did correctly identify the outliers, but it was not able to discover the sparse structure of the data as well as ROSPCA does.

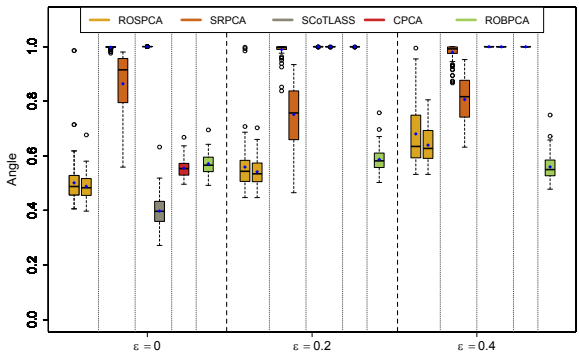


Figure 2.6: Angle values of ROSPCA, SRPCA, SCoTLASS, CPCA and ROBPCA at $p = 500$ and $\varepsilon = \{0, 0.2, 0.4\}$ for $n = 50$.

Consider now the high-dimensional simulations where $p = 500$. We now use the following grid of λ values: $\{0, 0.02, \ldots, 1.2\}$. For SRPCA with $n = 500$, we decreased the grid with λ values up to 0.6 instead of 1.2 to keep computations reasonable. Figures 2.6, 2.7 and 2.8 show the results for several sample sizes. As before, the bias and the dispersion of the angle becomes smaller when the sample size n increases. On uncontaminated data, the selection of λ is not successful for SCoTLASS (the BIC from Croux et al. (2013) returns even slightly worse results). However, the minimum angle boxplot shows that SCoTLASS can perform well, and ROSPCA attains similar performance to SCoTLASS's optimal performance in both boxplots. SRPCA shows very poor performance even when outliers are not present when λ is selected, and has worse results for the minimal angle values as well, indicating that intrinsically it may not be as accurate as SCoTLASS or ROSPCA. CPCA and ROBPCA have a comparable behaviour, which is inferior to the sparse methods. When contamination is introduced, SCoTLASS performs very poorly, as expected, while the optimal

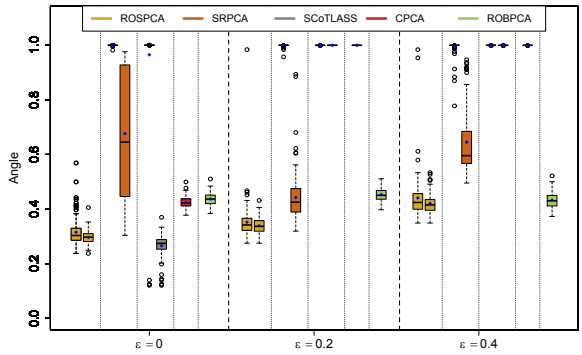


Figure 2.7: Angle values of ROSPCA, SRPCA, SCoTLASS, CPCA and ROBPCA at $p = 500$ and $\varepsilon = \{0, 0.2, 0.4\}$ for $n = 100$.

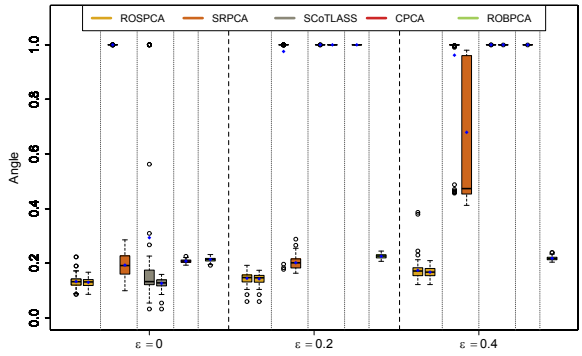


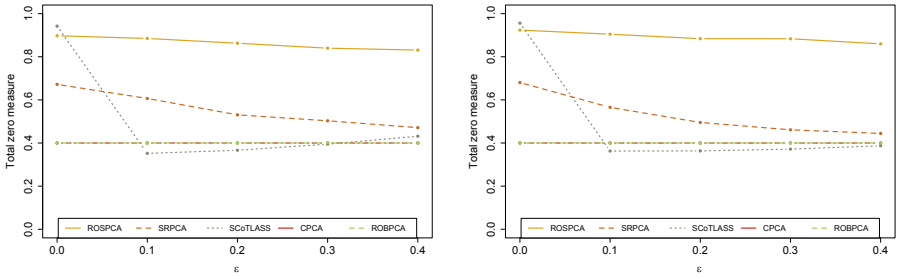
Figure 2.8: Angle values of ROSPCA, SRPCA, SCoTLASS, CPCA and ROBPCA at $p = 500$ and $\varepsilon = \{0, 0.2, 0.4\}$ for $n = 500$.

performance of SRPCA and ROSPCA is only slightly worse than when the data is not contaminated, and ROSPCA continues to show successful λ selection. When $\varepsilon = 0.4$, SRPCA does however show higher bias than for lower ε, unlike ROSPCA. CPCA is no longer reliable, whereas the performance of ROBPCA remains stable.

Sparsity

In addition to estimating a model that is not influenced by outliers, it is also important to estimate the correct sparsity. The *zero measure* is one way to compare how correctly each of the methods estimates the sparse $\boldsymbol{P}$. For each element of $\boldsymbol{P}$, it is equal to 1 if the estimated and true value are both zero or both non-zero, and 0 otherwise. We then take the average zero measure over all elements of $\boldsymbol{P}$ and all 500 simulations which we call the *total zero measure*. We need to specify when an element is "equal to zero" because it can be that an element of $\boldsymbol{P}$ is very small but different from zero. We say that all elements with an absolute value smaller than 10^{-5} are "equal to zero".

In Figures 2.9a and 2.9b, we see that ROSPCA accurately discerns the sparse structure of $\boldsymbol{P}$, even when $n = 50$ and $\varepsilon = 0.4$. SRPCA steadily demonstrates weaker performance as ε increases, whereas SCoTLASS performs well for $\varepsilon = 0$, and uniformly poorly for higher values of ε. The zero measure plots for larger sample sizes are very similar to the plot for $n = 100$. These results show that ROSPCA not only gives robust PCA estimates but is also better at detecting the sparse structure of the data. CPCA and ROBPCA hardly yield zero loading elements, so their zero measure is almost constantly equal to 40%, which is the percentage of non-zero entries in $\boldsymbol{P}$.



(a) $n = 50$ (b) $n = 100$

Figure 2.9: Total zero measure of ROSPCA, SRPCA, SCoTLASS, CPCA and ROBPCA for (a) $n = 50$ and (b) $n = 100$.

The zero measure is less useful in the high-dimensional setting because perfect sparsity for all zero loadings is more difficult to achieve. This results in zero measures that are comparatively more difficult to interpret than those shown in Figures 2.9a and 2.9b, since two methods may appear to give similar results by this measure, while a close inspection of the loadings reveals substantial differences.

The λ selection performance of ROSPCA, SRPCA and SCoTLASS

As explained in Section 2.2.6, we use the BIC-type criterion (2.6) to select the sparsity parameter λ of ROSPCA and SCoTLASS (since no criterion is proposed in Jolliffe et al. (2003)). For SRPCA, we use the BIC proposed by Croux et al. (2013). We looked at 101 (equidistant) values of λ over the interval in which complete sparsity is attained: $[0, 2.5]$, i.e. $\{0, 0.02, \ldots, 2.48, 2.5\}$. To provide insight into the role of robustness in this process, we introduce $\varepsilon = 20\%$ contamination. In Figure 2.10, we display the quantile plots of the angle values obtained by these methods over the 500 simulated datasets for $n = 100$ and $\varepsilon = 0.2$ as a function of λ. It depicts the median (solid lines) and first (dotted lines) and third quartile (dashed lines) of angle values for a given λ value over the 500 simulations.

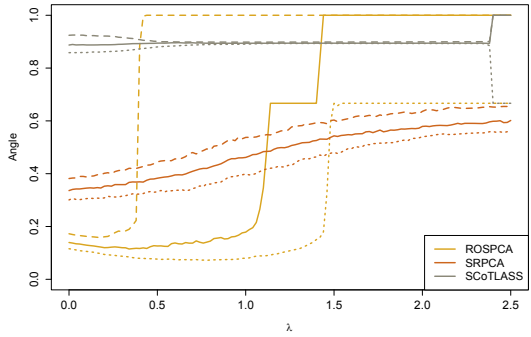


Figure 2.10: Quantile plots of the angle values for ROSPCA, SRPCA and SCoTLASS as a function of λ.

Examining the angle values corresponding to fits for each of the methods using different values of λ reveals a pattern correlated with the robustness of the methods. The angle values for SCoTLASS, tend to be fairly constant and high across the range of λ. This reflects the fact that the models are all influenced by outliers, and in comparison the sparsity of the model has very little impact of the angle. The quantile plot for SRPCA is not as flat as that of SCoTLASS and is considerably lower, but shows a steadily increasing angle value as λ is increased. Since this method is robust, it can attain decent fits with non-sparse models, but including sparsity makes it vulnerable to missing the outliers and finding a worse fit. This has the consequence that even though the true data is sparse, a full SRPCA model attains the lowest angle value since it allows for the most accurate outlier screening.

The quantile plot for SRPCA illustrates a trade-off between robustness and sparseness, where we find that contamination due to outliers tends to dominate the inaccuracy due to using a non-sparse model on sparse data (which is why the full SRPCA model has the lowest angle). The ROSPCA quantile plot shows that it is possible to account for both the sparse structure of the data and the outliers. For ROSPCA, the lowest value of λ (0 in our case) does not correspond to the lowest angle value. Rather, this is achieved by a sparse model, as we would expect. This is possible because ROSPCA has initially separated the outlier detection and sparsity steps before combining insights from both to return the final model. The first and third quantiles show that there is some variation in the angle values returned by ROSPCA for different values of λ, but the figures in Section 2.3.2 show that the value of λ selected by the BIC criterion is consistently close to the value of λ returning the minimal angle for each simulation.

2.4 Real data example

In this section we illustrate the behaviour of ROSPCA and SRPCA on the glass dataset introduced in Hubert et al. (2005). It consists of electron probe X-ray microanalysis (EPXMA) spectra over $p = 750$ wavelengths and 180 collected glass samples (Lemberge et al., 2000). Although the non-sparse ROBPCA performs well on this dataset, employing a sparse method may be interesting because when one consults the full loadings, the data actually appears to have a sparse structure. Figure 2.11 shows a heatmap of the absolute values of the centred data matrix where we used the componentwise median. We only plotted the wavelengths with numbers 120-400 because the rest of them are mostly non-informative (due to the sparse structure of the data). As noted in Hubert et al. (2005), two groups of outliers can be clearly identified in this dataset: the last 38 observations that were measured after the spectrometer was cleaned and calcium outliers with high values for two groups of wavelengths between 300 and 370.

With a robust sparse PCA analysis we hope to achieve outlier detection results comparable to ROBPCA while also obtaining sparse loadings that reflect the atomic structure of the glass samples. We do not standardise the data because all variables are expressed in the same units. The non-robustness of SCoTLASS means that it cannot reliably address the outliers present, so results are omitted.

Selecting the number of components to use in a sparse PCA model is more complicated than in classical PCA due to the inclusion of λ, which varies with k, but must also be selected. In Jolliffe et al. (2003), rather than providing a

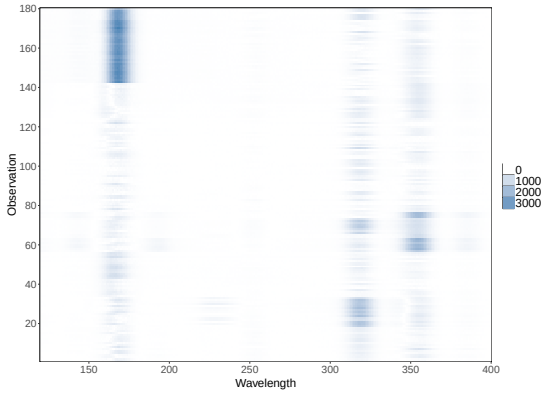


Figure 2.11: Glass data: heat map.

criterion for selecting k that accounts for sparsity, the authors apply the CPV criterion to a non-sparse PCA model. Then, they discuss the influence of a range of λ values over a model using that particular value of k. In Croux et al. (2013), the authors fit a robust, non-sparse PCA model with many components and then use those eigenvalues to select k for the sparse, robust model. Similarly, we use the eigenvalues of the robust, non-sparse PCA model described in Step 1 of ROSPCA. Since the SVD is computed on uncontaminated observations, we obtain eigenvalues for all possible $\min\{p, n-1\}$ components. We use the scree plot corresponding to these eigenvalues to select the number of components to retain, but automatic criteria such as the CPV can also be used.

The scree plot for ROSPCA (Figure 2.12) indicates that three or four components are sufficient to model the data well, and we select four components. Additionally we set the parameter $\alpha = 0.5$ to obtain maximal robustness. Hence $h_0 = \lceil 0.5 \times 180 \rceil + 1 = 91$. We also select $k = 4$ for SRPCA after consulting the scree plot for SRPCA with $\lambda = 0$.

Next, we perform the λ selection step for ROSPCA using our proposed BIC, and for SRPCA using the BIC of Croux et al. (2013). This yields λ values of 0.96 and 72.7, respectively. The running time for ROSPCA using $\lambda = 0.96$ was 146s, whereas SRPCA had a running time of 419s. For comparison we also include the ROBPCA results. As its scree plot is identical to that of ROSPCA (since the singular values are computed on the same subset of observations), we also use $k = 4$ components.

From the fitted models we can produce outlier maps showing the score distance and orthogonal distance of the observations in the dataset. We normalise these diagnostic plots by dividing each of the distances by its cut-off to make the

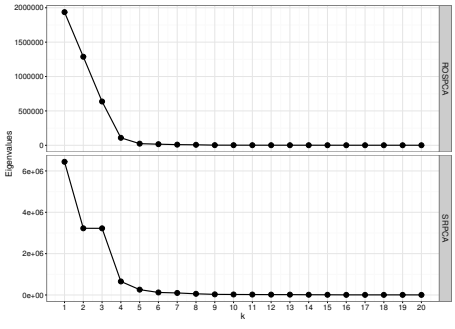


Figure 2.12: Glass data: scree plots for ROSPCA and SRPCA ($\lambda = 0$).

results visually comparable across methods. This gives us Figure 2.13. All three methods indicate the post-cleaning observations (orange) as bad leverage points, but SRPCA does not show the same discriminatory power as ROSPCA and ROBPCA. These two methods also clearly find several other orthogonal outliers and bad leverage points. This is useful for the practitioner because it provides a clear message that these observations warrant further investigation. Ignoring the boundary cases, we have indicated this set of outliers, as detected by ROSPCA, as open blue circles. Obviously ROBPCA identifies these outliers as well, but SRPCA rather declares them as ambiguous border cases with only larger score distances. Next, we compared the heatmap of the data in Figure 2.11 with these outlier maps, and noticed that almost all open blue circles correspond to calcium outliers which were highlighted on the heatmap. The three open blue circles that are close to the cut-off line for the score distances on the diagnostic plot of ROSPCA are however not clearly visible on the heatmap. Only a closer inspection of the raw data revealed that they are outlying on variables 215–245. Our robust multivariate analysis was able to detect this abnormal behaviour at once.

To study the sparsity, we plot the loadings of each of the methods in Figure 2.14 and tabulate the sparsity of each in Table 2.1. Unsurprisingly, ROBPCA produces the least sparse loadings, with only 13 variables with all loadings less than the threshold of 10^{-5}. Nonetheless, the loadings are instructive as they give a sense of the full structure of the data and where sparsity might be obtained. Specifically, three groups of wavelengths (155–185, 310–335, 336–370) are particularly relevant. SRPCA attains the greatest sparsity, but given the poor outlier detection performance, it is likely that as we saw in the simulation studies, the λ selection procedure has been influenced by contamination. The sensitivity of the λ selection step to outliers underscores the need for a highly robust method. ROSPCA obtains loadings similar to those of ROBPCA, but

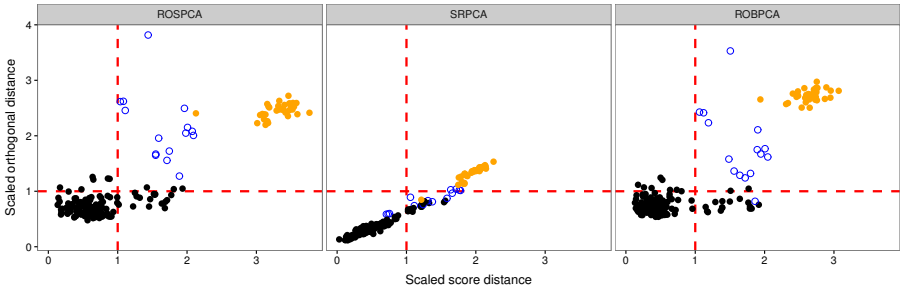


Figure 2.13: Glass data: scaled outlier maps of ROSPCA (with $\lambda = 0.96$), SRPCA ($\lambda = 72.7$) and ROBPCA. The orange points correspond to the measurements after the window has been cleaned. The open blue circles correspond to the other outliers identified by ROSPCA.

with the important distinction that loadings ROBPCA assigned small values to are now assigned no weight, resulting in 200 excluded variables. This increases the interpretability of the resulting model, while retaining accuracy. We note that a practitioner may choose a larger λ in an ad hoc way to further increase the sparsity of ROSPCA and that for a value of λ giving similar sparsity to that of SRPCA, ROSPCA still identifies the outliers correctly.

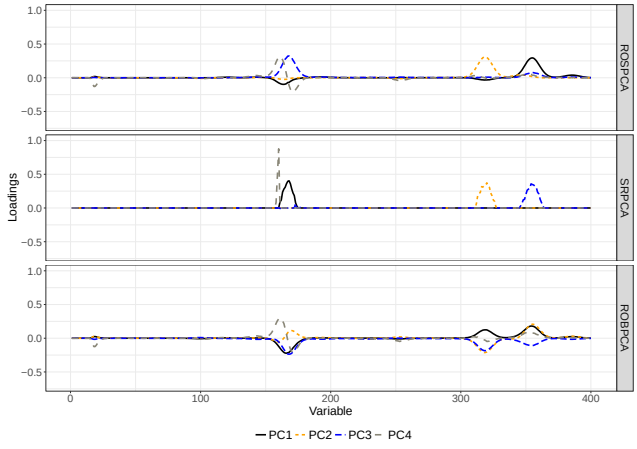


Figure 2.14: Glass data: loadings of ROSPCA (with $\lambda = 0.96$), SRPCA ($\lambda = 72.7$) and ROBPCA. Loadings on wavelengths with indices above 400 are small for all methods and are excluded from the plot.

	ROSPCA	SRPCA	ROBPCA
PC1	359	14	733
PC2	272	17	735
PC3	491	34	737
PC4	408	4	736
No. of excluded variables	200	696	13

Table 2.1: Glass data: number of non-zero loadings (larger than 10^{-5}) for each method per PC. The bottom row is the number of variables that have zero loadings (smaller than 10^{-5}) on all four PCs.

	ROSPCA-ROBPCA	ROBPCA-SRPCA	SRPCA-ROSPCA
Angle	0.040	0.731	0.725

Table 2.2: Glass data: angles between the obtained loadings using ROSPCA, ROBPCA and SRPCA.

Finally, we also compare the obtained loadings using the angle measure, results are shown in Table 2.2. We see that the ROSPCA and ROBPCA subspaces are similar and that the SRPCA subspace differs a lot from the other two subspaces. One could also visually deduce these conclusions from inspecting Figure 2.14.

The results for the glass dataset reinforce our findings from the simulations. Since the outliers are in two groups, we find that SRPCA does well at detecting the more obvious post-cleaning ones, but struggles to find the more nuanced calcium outliers. As in the simulations, ROSPCA both detects the outliers accurately and finds a plausible sparse structure.

2.5 Skewed data

All of the methods discussed so far work best when the non-outlying observations are symmetrically distributed. When the data is skewed, which is typically the case for financial returns data, they will tend to consider observations from the tail as outliers even though they may come from the distribution of the majority of the data. Under this setting, ROSPCA requires an adjustment to the outlyingness measure and the cut-offs of the robust distances to accurately infer whether an observation is an outlier or not. To do this, we follow the approach of Hubert et al. (2009). Firstly, the Stahel-Donoho outlyingness (2.5) in step 1 is replaced with an *adjusted outlyingness (AO)* measure based on the

adjusted boxplot (Hubert and Vandervieren, 2008). It is defined as (Brys et al., 2005; Van der Veeken and Hubert, 2008):

$$AO_i = \max_{\boldsymbol{v}\in\boldsymbol{B}} \frac{|\boldsymbol{x}_i'\boldsymbol{v} - \text{med}(\boldsymbol{x}_j'\boldsymbol{v})|}{(c_2(\boldsymbol{v}) - \text{med}(\boldsymbol{x}_j'\boldsymbol{v}))I(\{\boldsymbol{x}_i'\boldsymbol{v} > \text{med}(\boldsymbol{x}_j'\boldsymbol{v})\}) + (\text{med}(\boldsymbol{x}_j'\boldsymbol{v}) - c_1(\boldsymbol{v}))I(\{\boldsymbol{x}_i'\boldsymbol{v} < \text{med}(\boldsymbol{x}_j'\boldsymbol{v})\})} \tag{2.7}$$

where c_1 corresponds to the smallest observation which is greater than $Q_3 - 1.5\exp(3MC)\text{IQR}$ and c_2 corresponds to the largest observation which is smaller than $Q_3 + 1.5\exp(3MC)\text{IQR}$. Here, Q_1 and Q_3 are the first and third quartile of the projected data, $\text{IQR} = Q_3 - Q_1$ and MC is the medcouple (Brys et al., 2004), a robust measure of skewness. When $\text{MC} < 0$, we replace $\boldsymbol{v}$ by $-\boldsymbol{v}$. The denominator of the adjusted outlyingness makes sure that fewer non-outlying data points are incorrectly flagged as outliers at skewed distributions. Secondly, the cut-off value for the ODs is changed to the largest OD_i smaller than $Q_3(\text{OD}) + 1.5\exp(3\max\{\text{MC}(\text{OD}), 0\})\text{IQR}(\text{OD})$. Thirdly, the score distances are no longer computed using (2.2) but are obtained as the adjusted outlyingness measure applied to the scores. The cut-off value for the SDs is computed in the same manner as the one for the ODs. The cut-offs for the robust distances hence do no longer depend on quantiles of theoretical distributions as was the case for the non-adjusted versions of ROBPCA and ROSPCA. We denote the adjusted versions of ROBPCA and ROSPCA by *ROBPCA_AO* and *ROSPCA_AO*, respectively. Note that there is no straightforward adjustment of SRPCA to skewed data.

To show the performance of ROSPCA_AO, we look at the weekly log-returns of 30 stocks in the Dow Jones Industrial Average (DJI) between January 1991 and January 2001. This dataset is also used in Chapter 17 in Ruppert (2010). Based on the scree plot in Figure 2.15 we select $k = 3$ components. We now use $\alpha = 0.75$, hence $h_0 = \lceil 0.75 \times 505 \rceil + 1 = 390$. The choice of λ for ROSPCA and ROSPCA_AO is again based on the BIC-type criterion. In this case, the values of λ corresponding to the minimal BIC are close to zero for both methods which results in barely sparse models. Therefore, we opt to take a larger value of λ to get more sparsity while making sure that the increase in BIC is limited. This leads to choices $\lambda = 1.45$ for ROSPCA_AO and $\lambda = 1.35$ for ROSPCA.

In Figure 2.16, we plotted the scaled outlier maps of ROSPCA, ROSPCA_AO, ROBPCA and ROBPCA_AO. It is immediately clear that ROBPCA and ROSPCA flag too many observations as being outliers due to the skewed nature of the data. Instead, ROSPCA_AO and ROBPCA_AO only flag a few weeks as being outlying. The methods thus indicate that these weeks require further investigation since they deviate from all other weeks. Looking into these outliers we see that most of them are weeks in March and October 2000 where severe losses, or gains, occurred. Examples include:

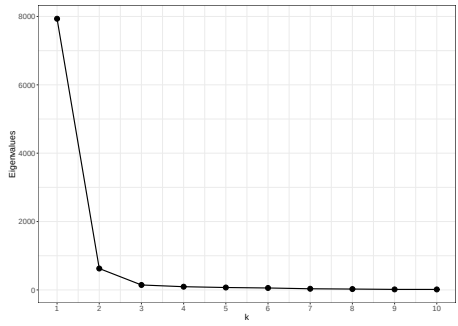


Figure 2.15: DJI data: scree plot for ROSPCA_AO ($\lambda = 0$).

- The week of 14 April 2000 when the DJI lost 5.66% (on a single day) in response to a US government report stating consumer prices were stronger than expected. This caused fears for inflation.
- The week of 12 October 2000 where the DJI fell more than 3% due to increasing oil prices which were caused by unrest in the Middle East (i.a. USS Cole bombing and the Rammallah incident in Israel).

Finally, we look at the plots of the loadings of ROSPCA_AO and ROBPCA_AO in Figure 2.17 to see the effect of the sparsity. Moreover, in Table 2.3 we give the number of excluded variables per PC, and the number of variables that is excluded on all three PCs. The first PC is determined by 22 variables, but the second and the third PC are only determined by 3 and 4 variables, respectively. The corresponding companies for PC2 are information technology companies:

- HP Inc. (HWP), Intel Co. (INTC) and Microsoft Co. (MSFT).

The third PC consists of:

- AT&T Inc. (T), Home Depot Inc. (HD), Wal-Mart Stores Inc. (WMT) and Walt Disney Co. (DIS).

Six variables do not contribute to any of the three PCs and are thus found to be unimportant according to the PCA analysis:

- Exxon Mobil Co. (XOM), Johnson & Johnson (JNJ), Altria Group Inc. (MO), Procter & Gamble Co. (PG) and SBC Communications Inc. (SBC).

In contrast, all variables contribute to the PCs for the ROBPCA_AO method. Moreover, it is not clear based on the loadings plot which companies are leading determinants for these PCs.

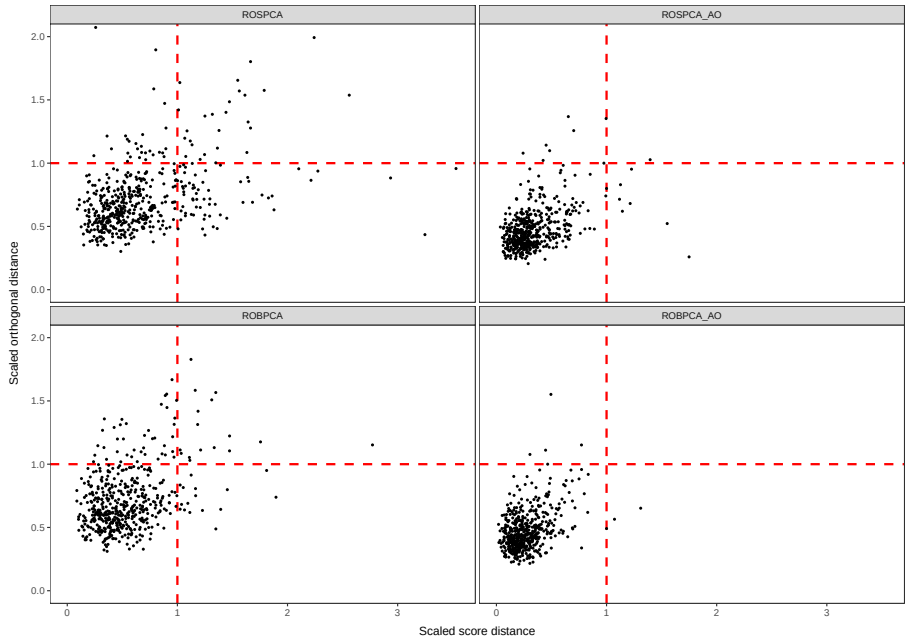


Figure 2.16: DJI data: scaled outlier maps of ROSPCA ($\lambda = 1.35$), ROSPCA_AO ($\lambda = 1.45$), ROBPCA and ROBPCA_AO.

	ROSPCA_AO	ROBPCA_AO
PC1	22	30
PC2	3	30
PC3	4	30
No. of excluded variables	5	0

Table 2.3: DJI data: number of non-zero loadings (larger than 10^{-5}) for each method per PC. The bottom row is the number of variables that have zero loadings (smaller than 10^{-5}) on all three PCs.

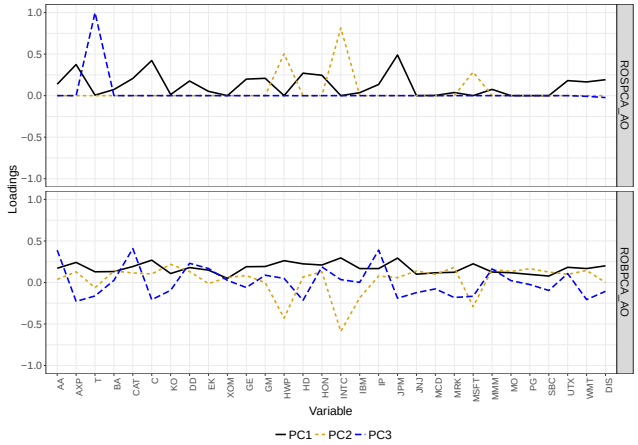


Figure 2.17: DJI data: loadings of ROSPCA_AO ($\lambda = 1.45$) and ROBPCA_AO.

2.6 Conclusions and research perspectives

We have detailed a new approach for robust sparse principal component analysis, ROSPCA, that is a modification of ROBPCA. Unlike existing methods for robust sparse PCA, ROSPCA prioritises the detection of the outliers rather than giving robustness and sparsity equal weight. Our results indicate that this approach is warranted. We observe that by first detecting and neutralising the outliers, ROSPCA is able to fit the sparse structure of the majority of the data with high accuracy. In comparisons with existing methods, we find that ROSPCA consistently obtains the best performance. Moreover, we proposed an adjusted version of ROSPCA that can handle skewed data.

In addition to good robustness and sparsity properties, ROSPCA is also computationally faster. One of the most important steps in performing a sparse PCA analysis is the selection of the λ parameter. A single execution of ROSPCA is faster than one of SRPCA, but this advantage is compounded when selecting λ since the robustness step only needs to be performed once.

The ROSPCA method is implemented in the `R` package *rospca* (Reynkens, 2017). Moreover, functions for the simulation study, an implementation of the BIC-type criterion (2.6) and the glass dataset are also included in the package. We also added an improved version of `PcaHubert` from *rrcov* (Todorov and Filzmoser, 2009) which uses fast implementations of the outlyingness measures (2.5) and (2.7) from the *mrfDepth* package (Segaert et al., 2017).

An interesting application of robust sparse PCA can be found in Plevka et al. (2016). They investigate travel behaviour determinants, and typically classical PCA is used in this context. Their analysis shows that outliers play a critical role in travel behaviour analysis, and that interpretation of the results is not straightforward. Therefore, they apply ROSPCA to travel behaviour data from Ghent. This reveals that variables associated with travel constraints (kids, luggage) are important determinants of travel behaviour.

Another robust sparse PCA method has recently been proposed by Greco and Farcomeni (2016). They introduce a robust covariance estimator in the SPCA method of Zou et al. (2006). Greco and Farcomeni (2016) add a comparison with ROSPCA in their simulations, but it is unclear how they selected λ for ROSPCA, and they considered only 20% of contamination. This makes a fair comparison between the new method and ROSPCA difficult.

This work opens the door to the development of robust sparse methods for high-dimensional data, such as sparse robust discriminant analysis and sparse partial least squares regression. Extensions of the ROBPCA-based methods, as in Vanden Branden and Hubert (2005) and Hubert and Vanden Branden (2003) can be studied. A theoretical study of the influence function of ROSPCA, extending the results of Debruyne and Hubert (2009), is also an interesting challenge for future research.

The DJI example indicates that the BIC-type criterion might select too low λ-values for skewed data. Further simulations are needed to investigate the behaviour of this criterion when selecting λ-values for ROSPCA_AO.

Rousseeuw et al. (2016) propose the directional outlyingness which is new measure of outlyingness for skewed distributions. This measure has a more robust scale (denominator) than the adjusted outlyingness (2.7), and can be computed using less operations. An interesting topic of further research is then to replace the adjusted outlyingness by the directional outlyingness in the adjusted versions of ROBPCA and ROSPCA.

Part II

Extreme Value Theory in Finance and Insurance

Chapter 3

Hunting for Black Swans in the European banking sector using extreme value analysis

This chapter is based on

Beirlant, J., Schoutens, W., De Spiegeleer, J., Reynkens, T. and Herrmann, K. (2016). Hunting for Black Swans in the European Banking Sector Using Extreme Value Analysis. In: J. Kallsen and A. Papapantoleon (eds.), *Advanced Modelling in Mathematical Finance: In Honour of Ernst Eberlein*, Springer International Publishing, Switzerland, pp. 147–166.

3.1 Introduction

Clearly, the recent financial crisis that started in 2007 can be used as a motivating example necessitating the use of extreme value analysis (EVA) in financial statistics. Bollerslev and Todorov (2011) studied the effect of the crisis for the S&P500 considering the contrast between the statistical probability measure and the risk neutral measure. Here, we study the tail behaviour of the negative log-returns of the weekly closing prices of listed stocks. Using techniques from extreme value methodology we propose to analyse the tail behaviour of a bank over two specific horizons:

- Pre-Crisis: from 1 January 1994 until 7 August 2007 (often referred to as the official starting date of the credit crunch crisis);
- Post-Crisis: from 8 August 2007 until 23 September 2014 (the cut-off date of our study).

More specifically, we will investigate how one could decide if the recent financial crisis was a Black Swan event for a given bank based on statistical differences between both sets of return data. We illustrate this approach using data from Barclays and Credit Suisse, two major European banks. Of course one should also connect such a statistical finding with economic indicators of a bank, whether it experienced a Black Swan event from an EVA perspective or not.

We restrict ourselves here to weekly return data. Indeed, financial return series may suffer from serial dependence such as volatility clustering, which violates the classical assumption of independence. Such serial dependence is at least much weaker in weekly returns. Using results from Hsing (1991) our statistical tests, however, will take serial dependence into account. In Figure 3.1 the negative weekly log-returns are plotted against time for the selected banks. The vertical scales are identical allowing to appreciate the impact of the crisis on the weekly losses for the different banks. We also add a vertical line indicating 7 August 2007.

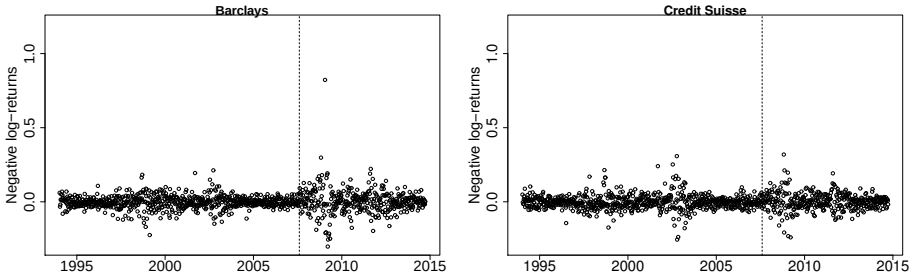


Figure 3.1: Negative weekly log-returns for Barclays and Credit Suisse.

EVA is designed for estimating extreme quantities of a statistical variable, such as the Value-at-Risk (VaR), which has become a popular risk measure. The models underlying EVA contain scale and shape parameters, and the statistical methods on which estimation of the scale and shape has been built, offer tools that can be used for general statistical inference such as the definition of appropriate tail models for a distribution at hand. Generalised autoregressive conditional heteroskedasticity (GARCH) models constitute a popular approach in analysing financial time series which exhibit volatility clustering. Here, however, we follow the approach outlined in Sun and Zhou (2014), using the

results from Hsing (1991) that Hill's (1975) estimator is still consistent for certain types of dependent data, such as GARCH processes. Moreover, for a bank which was badly hit by the crisis, the fitted shape parameter can lead to a near integrated-GARCH situation, which entails an inappropriate GARCH fit and unreliable estimates of the GARCH innovations.

Here, we look for indicators for truly significant changes in the log-returns through statistical tests for changes in scale and shape parameters, and by calculating the return period of the largest post-crisis loss, in view of the data before the crisis. We emphasise the use of graphical methods that support decision making. In the next section we recall the most important facts from EVA, and review the graphical and estimation methods along the above specifications. Next, we propose estimators for the scale parameter in case of Pareto-type distributions and provide some new asymptotic results. In Section 3.4 we go into the problem of threshold selection when performing inference on the shape and scale parameters. In particular we stress the use of bias reduction techniques which helps to come around the problem of choosing a particular threshold when performing statistical inferences on the parameters. In the final section of this chapter we make the link with economic indicators.

3.2 A recollection from univariate extreme value methodology

3.2.1 Max-domain of attraction

We briefly recollect some facts from EVA. Recent books that have appeared on the subject provide more details: Embrechts et al. (1997), Coles (2001), Beirlant et al. (2004), Castillo et al. (2005), de Haan and Ferreira (2006), Reiss and Thomas (2007) and Resnick (2007). Beirlant et al. (2005) give an overview of EVA and apply it in a financial risk context. EVA is based on the limit result for normalised partial maxima of i.i.d. random variables $X_1, \ldots, X_n$. Let $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$ denote the ordered observations and hence $X_{n,n} = \max\{X_1, \ldots, X_n\}$. The limit theorem (Fisher and Tippett, 1928; Gnedenko, 1943) is then formulated as follows: if there exist normalising constants $a_n > 0$ and b_n such that for all x,

$$\lim_{n\to\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = G(x), \tag{3.1}$$

for some non-degenerate distribution function G, then G is necessarily of extreme value type; that is, up to an affine change of variables, one has

$$G(x) = G_\xi(x) = \exp\left(-(1+\xi x)^{-1/\xi}\right) \text{ if } x > -1/\xi \tag{3.2}$$

for some real value ξ. The parameter ξ is termed the extreme value index (EVI), which is of prime interest in EVA. When $\xi = 0$, $G_0(x)$ is to be read as $\exp(-\exp(-x))$. If (3.1) holds, we say that the distribution, which underlies the data $X_1, X_2, \ldots$, is in the max-domain of attraction (MDA) of G_ξ. The limiting distribution functions in (3.1) are then max-stable. They are indeed the unique max-stable laws.

The EVI ξ governs the behaviour of the right-tail of the distribution. The Fréchet domain of attraction ($\xi > 0$) contains heavy-tailed distributions like the Pareto and the Student t-distributions, i.e. tails of a negative polynomial type and infinite right endpoint. Short-tailed distributions, with a finite right endpoint like the beta distributions, belong to the Weibull MDA with $\xi < 0$. Finally, the Gumbel MDA corresponding to $\xi = 0$ contains a great variety of distributions with an exponentially decreasing tail, such as the exponential, the normal and the gamma distributions, but not necessarily with an infinite right endpoint.

In order to characterise the MDAs in a mathematically correct way, there are now two possibilities: model descriptions through the cumulative distribution function (CDF) $F(x) = P(X \leq x)$ (*probability view*) or through the quantile function Q, defined as the inverse function of F (*quantile view*).

Firstly, one can describe the MDAs through the stochastic behaviour of the so-called peaks over threshold (POT) $X - t$ given that $X > t$. the Pickands–Balkema–de Haan theorem (Pickands III, 1975; Balkema and de Haan, 1974) states that X is in the MDA of G_ξ if and only if for some sequence $\sigma_t > 0$ the conditional distribution of the scaled excesses as $t \to Q(1)$ converges to the generalised Pareto distribution (GPD)

$$P\left(\frac{X-t}{\sigma_t} \leq x \,\middle|\, X > t\right) \to H_\xi(x) = 1 - (1+\xi x)^{-1/\xi} \tag{3.3}$$

with $1 + \xi x > 0$ and $x > 0$. Remark that in case $\xi = 0$ the GPD is nothing else than the exponential distribution with distribution function $1 - \exp(-x)$ for $x > 0$.
From this, one chooses an appropriate threshold t and hopes for a reasonable rate of convergence in (3.3). Fitting the GPD with survival function $\left(1 + \frac{\xi}{\sigma}x\right)^{-1/\xi}$ to the excesses $X_i - t$ for those data X_i for which $X_i > t$, one estimates the shape parameter ξ and the scale σ for instance by maximum likelihood. In

practice t can be chosen as one of the largest data, e.g. the $(k+1)$th largest data point $X_{n-k,n}$, for some $1 < k < n$.

Secondly, through the work of de Haan (1970, 1984) the MDA characterisation was constructed on the basis of the regular varying behaviour of the tail quantile function U, which is associated with the quantile function Q by $U(x) := Q(1-\frac{1}{x})$. The MDAs can indeed be characterised by the extended regular variation property specifying the difference between high quantiles corresponding to tail proportions that differ by $100x\%$:

$$F \in MDA(\xi) \iff \lim_{u\to\infty} \frac{U(ux)-U(u)}{a(u)} = \frac{x^\xi - 1}{\xi} \tag{3.4}$$

for every real valued x and some positive function a, and where the expression on the right equals $\ln x$ for $\xi = 0$.

In the specific case of the Fréchet MDA with $\xi > 0$, the extended regular variation property (3.4) corresponds to regular variation of U with index $\xi > 0$:

$$F \in MDA(\xi > 0) \iff U(x) = x^\xi \ell(x), \tag{3.5}$$

where ℓ is a slowly varying function defined by $\lim_{u\to\infty} \frac{\ell(ux)}{\ell(u)} = 1$, for all $x > 0$. Then, condition (3.3) specifies the regular variation of the right tail function $\bar{F} := 1 - F$ with index $1/\xi$. The elements of this MDA are termed Pareto-type distributions. Remark that the regular variation of $\bar{F}$ is equivalent to stating that as $t \to \infty$

$$P\left(\frac{X}{t} > x \,\middle|\, X > t\right) \to x^{-1/\xi}, \;\; x > 1, \tag{3.6}$$

which then forms a simplified POT approach in comparison with (3.3).
Almost all authors consider the following subclass of Pareto-type distributions, which was first introduced in Hall (1982):

$$\bar{F}(x) = Ax^{-1/\xi}\left(1 + bx^{-\beta}(1+o(1))\right), \tag{3.7}$$

$$U(x) = A^\xi x^\xi \left(1 + \xi b A^{-\xi\beta} x^{-\xi\beta}(1+o(1))\right), \quad \text{as } x \to \infty, \tag{3.8}$$

where $A > 0$ is then the scale parameter, while $\beta > 0$ and b are the second-order shape and scale parameters. This extra assumption then allows to derive specific approximations for the bias and variance of the estimators, and to derive bias reduced estimators as discussed below.

3.2.2 Estimation when $\xi > 0$

Assumption (3.5) can be graphically verified using log-log plots, i.e. Pareto quantile-quantile (QQ)-plots,

$$\left(\ln \frac{n+1}{i}, \ln X_{n-i+1,n}\right), \quad i = 1, \ldots, n, \tag{3.9}$$

which, for some k, should then be ultimately linear for a set of largest values $X_{n-k,n} \leq X_{n-k+1,n} \leq \ldots \leq X_{n,n}$. The classical Hill (1975) estimator $H_{k,n}$ of $\xi > 0$

$$H_{k,n} = \frac{1}{k} \sum_{j=1}^{k} \ln X_{n-j+1,n} - \ln X_{n-k,n}, \tag{3.10}$$

can be motivated as an estimator of the slope of the least squares regression line based on the final k points in the Pareto QQ-plot and passing through an appropriately chosen anchor point $\left(\ln \frac{n+1}{k+1}, \ln X_{n-k,n}\right)$, see Beirlant et al. (1996). It can also be derived as a maximum likelihood estimator of ξ using the simple Pareto model in the right hand side of (3.6) based on the relative excesses $\frac{X_{n-j+1,n}}{X_{n-k,n}}$, $j = 1, \ldots, k$, over the random threshold $X_{n-k,n}$.

Hsing (1991) derived the asymptotic distribution of $H_{k,n}$ for weakly dependent series. Under (3.7), as $k, n \to \infty$ and $k/n \to 0$, this leads to

$$\sqrt{k}\left(H_{k,n} - \xi - \frac{B(n/k)}{1+\xi\beta}\right) \to_d \mathcal{N}\left(0, \xi^2(1 + \chi + \omega - 2\psi)\right) \tag{3.11}$$

as $n \to \infty$ and $k/n \to 0$, where $B(n/k) = -\xi\beta b A^{-\xi\beta}(k/n)^{\xi\beta}$, and χ, ω, ψ are parameters of serial dependence, being 0 in case of independence. Under the condition $\sqrt{k}B(n/k) \to \lambda$ as $k, n \to \infty$ and $k/n \to 0$ we then obtain

$$\sqrt{k}\left(H_{k,n} - \xi\right) \to_d \mathcal{N}\left(\frac{\lambda}{1+\xi\beta}, \xi^2(1 + \chi + \omega - 2\psi)\right).$$

Estimators $\hat{\chi}, \hat{\omega}, \hat{\psi}$ are given in (3.6) in Hsing (1991). Furthermore, Sun and Zhou (2014) showed that a GARCH(1,1) dependence structure fits to the approach of Hsing (1991).

From (3.11) it follows that $H_{k,n}$ can have high bias for a large range of k values. This bias originates from the fact that the estimators are based on (3.6) replacing the limit by an equality, which is inaccurate for too large values of k. Theoretically, this k-region is represented by $\sqrt{k}B(n/k) \to \lambda > 0$ as $k, n \to \infty$, $k/n \to 0$. Accommodation of bias has been considered recently in a number of papers in case of i.i.d. data. Bias reduced estimators typically

exhibit plots which are more horizontal as a function of k. In case the tail under consideration is a composition of two different Pareto components, the corresponding levels of the estimates are better visible. In that sense such estimators are useful as a diagnostic tool in order to interpret Hill plots $(k, H_{k,n})$ and plots of other tail estimators. For instance, choosing a value of k as large as possible, with the original and bias reduced version of the estimator approximately equal, leads to an estimate with a smaller bias and a variance as small as possible. Along the probability view, bias reduction can be obtained by replacing the Pareto fit in (3.6) by an extended Pareto distribution (EPD) with distribution function

$$G_{\xi,\kappa,\beta}(y) = 1 - \left(y(1 + \kappa(1 - y^{-\beta}))\right)^{-1/\xi}, \; y > 1,$$

to the relative excesses $\frac{X_{n-j+1,n}}{X_{n-k,n}}$, $j = 1, \ldots, k$ using maximum likelihood (see Beirlant et al., 2009). This EPD approximation follows when approximating the left hand side of (3.6) under (3.7) with $\kappa = \kappa_t = \xi b t^{-\beta}$.

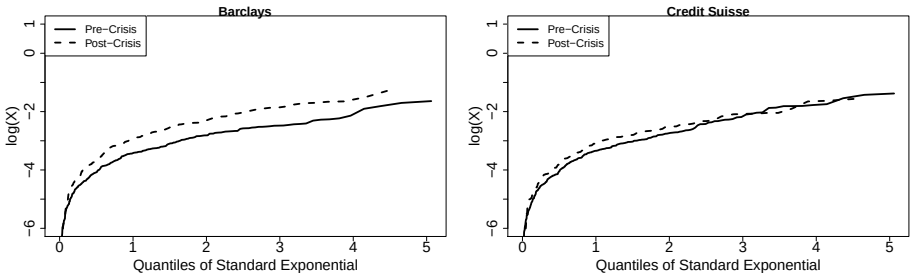


Figure 3.2: Pareto QQ-plots for the pre- (solid line) and post-crisis (dashed line) negative log-returns for Barclays and Credit Suisse.

In Figure 3.2, we gather the Pareto QQ-plots. The Hill plots with the original $H_{k,n}$ and using the EPD approximation with (ξ, κ) estimated by maximum likelihood per k are shown in Figure 3.3. We set $\rho = -\beta\xi$ equal to -1 and hence $\hat{\beta} = -\rho/\hat{\xi} = 1/\hat{\xi}$. Because of the different sample sizes for pre- and post-crisis data we plot the estimates against the ratio k/n. The Pareto QQ-plots show that there is barely any difference in slope between the pre- and post-crisis data. The plots for the shape estimators seem to confirm this. For Barclays, k/n values around 0.1 seem to be suitable along the abovementioned guideline, since for both periods, the Hill and EPD estimates remain rather close for $k/n \leq 0.1$, in contrast to the larger k values. In the case of Credit Suisse, the ultimate top portion of the Pareto QQ-plot appears to be concave leading to decreasing Hill estimates as k decreases, meeting the bias reduced estimator only at the smallest k values, say up to $k/n \approx 0.02$. For both banks, the Hill estimates for the pre- and post-crisis are close in the suitable region for

k/n. However, in case of Barclays, the Pareto QQ-plot of the post-crisis data lies higher than the one of the pre-crisis data, indicating a change in scale since it follows from (3.8) that the Pareto QQ-plot is an approximation of the graph $\left(\ln \frac{n+1}{i}, \ln U(\frac{n+1}{i})\right) = \left(\ln \frac{n+1}{i}, \xi \ln A + \xi \ln \frac{n+1}{i}\right)$ for $i = 1, \ldots, n$. In the Credit-Suisse case, both Pareto QQ-plots are close and hence no change in scale can be deduced.

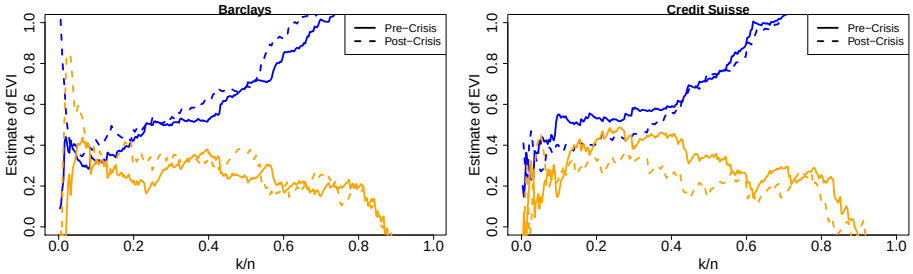


Figure 3.3: Hill (blue) and EPD (orange) estimates as a function of k/n for the pre- (solid line) and post-crisis (dashed line) negative log-returns for Barclays and Credit Suisse.

3.3 Estimating the scale parameter

Following the suggestion made in Einmahl et al. (2016) one can also inspect for changes in the scale parameter A introduced in (3.7)-(3.8). An initial estimator for A is given by

$$\hat{A}_{k,n} = \frac{k+1}{n+1} X_{n-k,n}^{1/H_{k,n}}. \tag{3.12}$$

The following theorem provides an asymptotic normality result for this estimator which is valid for dependent data.

Theorem 3.1. *Under the conditions of Theorem 3.3 in Hsing (1991) and under* (3.7)*, when* $k, n \to \infty$*,* $k/n \to 0$ *and* $\sqrt{k}B(n/k) \to \lambda$*, we have that*

$$\frac{\sqrt{k}\xi}{\ln U(n/k)} \left(\frac{\hat{A}_{k,n}}{A} - 1 \right) \to_d \mathcal{N}\left(\frac{-\lambda}{1+\xi\beta}, 1 + \chi + \omega - 2\psi \right).$$

Theorem 3.1 shows that the scale estimator can have large bias. Using $\hat{\xi}_{k,n}$ and $\hat{\kappa}_{k,n}$, the EPD estimators of ξ and κ, we get the following bias reduced estimator of A:

$$\hat{A}_{k,n}^{EP} = \frac{k+1}{n+1} X_{n-k,n}^{1/\hat{\xi}_{k,n}} \left(1 - \frac{\hat{\kappa}_{k,n}}{\hat{\xi}_{k,n}} \right). \tag{3.13}$$

We provide an intuitive derivation for both scale estimators in Appendix A.1. Theorem 3.2 gives the asymptotic distribution of $\hat{A}_{k,n}^{EP}$ in case of independent data. It is then clear that $\hat{A}_{k,n}^{EP}$ is indeed a bias reduced estimator of A. The proofs of both theorems are postponed to Appendix A.2.

Theorem 3.2. *Assuming $X_1, \ldots, X_n$ are independent and identically distributed following* (3.7)*, when $k, n \to \infty$, $k/n \to 0$ and $\sqrt{k}B(n/k) \to \lambda$, we have that*

$$\frac{\sqrt{k}\xi}{\ln U(n/k)} \left(\frac{A}{\hat{A}_{k,n}^{EP}} - 1 \right) \to_d \mathcal{N}\left(0, \left(\frac{1+\xi\beta}{\xi\beta}\right)^2\right).$$

In Figure 3.4, $\ln \hat{A}_{k,n}$ and $\ln \hat{A}_{k,n}^{EP}$ are plotted for the two selected banks with the pre- and post-crisis series. We can again select suitable regions based on the closeness of the scale estimator and the bias reduced version. We then choose $k/n \approx 0.1$ for Barclays and $k/n \approx 0.02$ for Credit Suisse. We see that there is some difference in scale estimates between the pre- and post-crisis data for Barclays while much less for Credit Suisse (for these values of k/n).

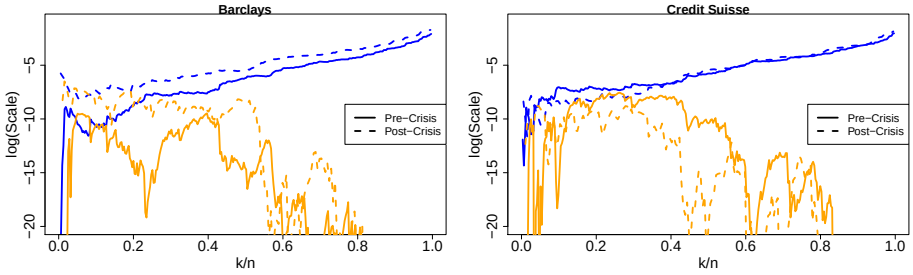


Figure 3.4: Scale estimates $\hat{A}_{k,n}$ (blue) and bias reduced scale estimates $\hat{A}_{k,n}^{EP}$ (orange), in log-scale, as a function of k/n for the pre- (solid line) and post-crisis (dashed line) negative log-returns for Barclays and Credit Suisse.

3.4 Testing for Black Swans

We define a Black Swan as a highly improbable event with large consequences. Therefore, as a first indicator we consider the probability of obtaining a loss at least as big as the largest loss post-crisis, *in view of the data information before the crisis*. We express this in terms of the corresponding return period. Secondly, we test for significant differences in scale and shape parameters between pre- and post-crisis periods. While it is difficult to define a Black Swan through a

minimal return period and/or a maximal P-value level, we will argue that the financial crisis can be considered as a Black Swan in the Barclays case, while it is not in the Credit Suisse case.

3.4.1 Return periods of worst negative log-returns

Here, and in the sequel, we denote the number of pre-crisis, respectively post-crisis, negative log-returns by n_1, respectively n_2, and the ordered pre-crisis, respectively post-crisis, negative log-returns by $x_{1,n_1}, \ldots, x_{n_1,n_1}$, respectively $y_{1,n_2}, \ldots, y_{n_2,n_2}$. We also use the superscripts (X) and (Y) to indicate the pre-crisis, respectively, post-crisis data. The return period can now be denoted by $r_{\max} = 1/P(X > y_{n_2,n_2})$. Then, applying the Weissman (1978) estimator following from the approximation (3.6),

$$\hat{r}_{\max,k} = \frac{1}{\hat{P}_k(X > y_{n_2,n_2})} = \frac{n_1+1}{k+1}\left(\frac{y_{n_2,n_2}}{x_{n_1-k,n_1}}\right)^{1/H^{(X)}_{k,n_1}}, \; k = 1, \ldots, n_1.$$

In a similar way as in the proof of Theorem 3.1 in Appendix A.2 (see also Theorem 4.4.7 in de Haan and Ferreira, 2006), one can show that, treating y_{n_2,n_2} as a fixed number,

$$\frac{\sqrt{k}}{\sqrt{1+\ln^2\left(\frac{k}{n_1}r_{\max}\right)}}\left(\ln r_{\max} - \ln \hat{r}_{\max,k}\right)$$

is asymptotically normal with asymptotic variance $1 + \chi + \omega - 2\psi$. Hence an approximate 95% asymptotic lower confidence bound for $\ln r_{\max}$ is given by

$$\ln \hat{r}_{\max,k} - \frac{1.645}{\sqrt{k}}\sqrt{1+\ln^2\left(\frac{k}{n_1}\hat{r}_{\max,k}\right)}\sqrt{1+\hat{\chi}+\hat{\omega}-2\hat{\psi}}. \tag{3.14}$$

As described in Beirlant et al. (2009), a bias reduced version for return periods can be constructed by replacing the simple Pareto distribution by the EPD in the right hand side of (3.6):

$$\hat{r}^{EP}_{\max,k} = \frac{n_1+1}{k+1}\left(1 - G_{\hat{\xi}_{k,n_1},\hat{\kappa}_{k,n_1},\hat{\beta}_{k,n_1}}\left(\frac{y_{n_2,n_2}}{x_{n_1-k,n_1}}\right)\right)^{-1}, \; k = 1, \ldots, n_1.$$

In Figure 3.5 we plot $\ln \hat{r}_{\max,k}$ and $\ln \hat{r}^{EP}_{\max,k}$ as a function of k/n_1 for the two selected banks, jointly with the lower bounds (3.14) (dashed lines). Choosing $k/n_1 \approx 0.1$ where the different estimators coincide, we obtain for Barclays a return period $e^{10} \approx 22000$ weeks. This return period corresponds to $2\times 423 = 846$

years using an equal frequency for negative and positive log-returns. This is in sharp contrast with the corresponding return period $e^6 \approx 400$ weeks or $2 \times 7.7 = 15.4$ years for Credit Suisse.

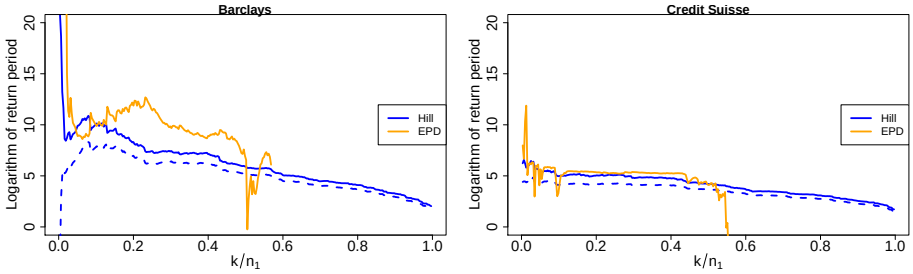


Figure 3.5: Estimates of the return periods for obtaining a weekly loss as big as the largest loss post-crisis in view of the data information before the crisis: $\ln \hat{r}_{\max,k}$ (blue) and $\ln \hat{r}^{EP}_{\max,k}$ (orange), as a function of k/n_1, for Barclays and Credit Suisse. Approximate 95% asymptotic lower confidence bounds for $\ln \hat{r}_{\max}$ are shown by the dashed lines.

3.4.2 Testing for differences in shape or scale

We now want to test more formally if there is a significant difference in at least the shape or the scale parameter. We consider the $\alpha = 5\%$ significance level.

In order to test $H_0^{(\xi)} : \xi^{(X)} \geq \xi^{(Y)}$ versus $H_1^{(\xi)} : \xi^{(X)} < \xi^{(Y)}$ we can use the test statistic

$$T^{(\xi)}_{k_1,k_2,n_1,n_2} = \frac{H^{(Y)}_{k_2,n_2} - H^{(X)}_{k_1,n_1}}{\sqrt{\frac{(H^{(Y)}_{k_2,n_2})^2\,(1+\hat{\chi}_2+\hat{\omega}_2-2\hat{\psi}_2)}{k_2} + \frac{(H^{(X)}_{k_1,n_1})^2\,(1+\hat{\chi}_1+\hat{\omega}_1-2\hat{\psi}_1)}{k_1}}}$$

with k_1 and k_2 appropriately selected number of extremes for pre- and post-crisis data, and $\hat{\chi}_1, \hat{\omega}_1, \hat{\psi}_1$ and $\hat{\chi}_2, \hat{\omega}_2, \hat{\psi}_2$ are the corresponding estimates for χ, ω, ψ for the pre- and post-crisis period respectively. Under equality of the tail indices the asymptotic distribution of $T^{(\xi)}_{k_1,k_2,n_1,n_2}$ is then standard normal for small values of k_1, k_2 such that $\sqrt{k_1}B^{(X)}(n_1/k_1) \to 0$ and $\sqrt{k_2}B^{(Y)}(n_2/k_2) \to 0$. Similarly, to test $H_0^{(A)} : A^{(X)} \geq A^{(Y)}$ versus $H_1^{(A)} : A^{(X)} < A^{(Y)}$ we use

$$T^{(A)}_{k_1,k_2,n_1,n_2} = \frac{\ln \hat{A}^{(Y)}_{k_2,n_2} - \ln \hat{A}^{(X)}_{k_1,n_1}}{\sqrt{\frac{\ln^2(n_2/k_2)\,(1+\hat{\chi}_2+\hat{\omega}_2-2\hat{\psi}_2)}{k_2} + \frac{\ln^2(n_1/k_1)\,(1+\hat{\chi}_1+\hat{\omega}_1-2\hat{\psi}_1)}{k_1}}}$$

which also follows asymptotically a standard normal distribution under equality in $H_0^{(A)}$.

As shown in Appendix A.3, the two tests are not independent. We therefore have to be prudent drawing conclusions. The joint test combines information of the two separate tests and uses the following hypotheses: $H_0 : H_0^{(\xi)} \cap H_0^{(A)}$ versus $H_1 : H_1^{(\xi)} \cup H_1^{(A)}$. From (A.2) in Appendix A.3 it follows that the determinant of the covariance matrix is asymptotically 0 and hence a bivariate Hotelling T^2 test cannot be performed. It is critical to control the probability for a type I error of the joint test, hence asking even more statistical evidence before concluding a Black Swan event. Using the Bonferroni correction we obtain that the probability for a type I error for the joint test is smaller than $\alpha = 5\%$ when using the $\alpha/2 = 2.5\%$ significance level for each test separately.

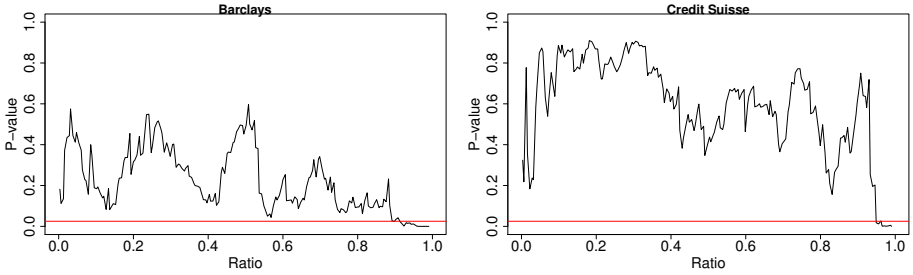


Figure 3.6: P-values for testing differences in shape using $T^{(\xi)}_{k_1,k_2,n_1,n_2}$ as a function of the ratio $k_1/n_1 = k_2/n_2$ for pre- and post-crisis negative log-returns for Barclays and Credit Suisse.

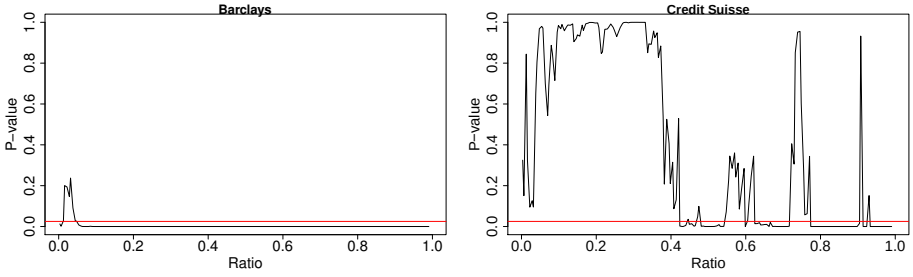


Figure 3.7: P-values for testing differences in scale using $T^{(A)}_{k_1,k_2,n_1,n_2}$ as a function of the ratio $k_1/n_1 = k_2/n_2$ for pre- and post-crisis negative log-returns for Barclays and Credit Suisse.

In Figures 3.6 and 3.7 we plot the P-values of the two asymptotic tests for equality of shape and scale against k/n under equality of the ratios $k_1/n_1 = k_2/n_2$. The red lines show the 2.5% significance level. From the discussion following Figures 3.3 and 3.4, using ratios around 0.1 for Barclays and around 0.02 for Credit Suisse corresponds to the lowest bias. The shape parameters do not show significant differences. The scale parameters show significant results for Barclays except for $k/n \leq 0.05$, whereas for Credit Suisse the scale parameters show strongly non-significant results for k/n smaller than 0.4. We now consider the P-values for testing scale differences for all possible choices of k_1, k_2 for Barclays and Credit Suisse. The 3-dimensional plots showing the P-values can be found in Figure A.1 and A.2 in Appendix A.4 where a red plane indicates the 2.5% significance level. Here, we consider the indicator function which takes value 1 when the P-value is below 2.5% and 0 otherwise. This function is plotted in Figure 3.8 and 3.9 where (light) blue and red correspond to 0 and 1, respectively, and the black dashed line indicates $k_1/n_1 = k_2/n_2$. In case of Barclays, the test for the difference in scale is non-significant only for large values of k_1, while in case of Credit Suisse non-significance also appears for small values of k_1 and k_2 together.

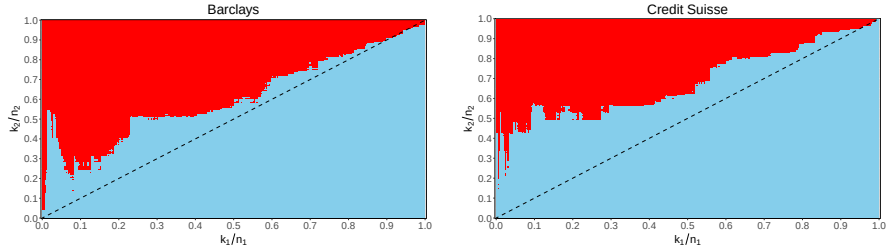


Figure 3.8: Indicator function for the event "P-value for the test using $T^{(\xi)}_{k_1,k_2,n_1,n_2}$ is below $\alpha/2 = 2.5\%$" for all possible choices of k_1 and k_2 for pre- and post-crisis negative log-returns for Barclays and Credit Suisse.

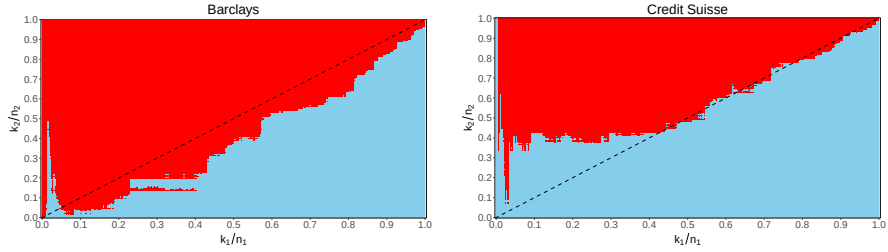


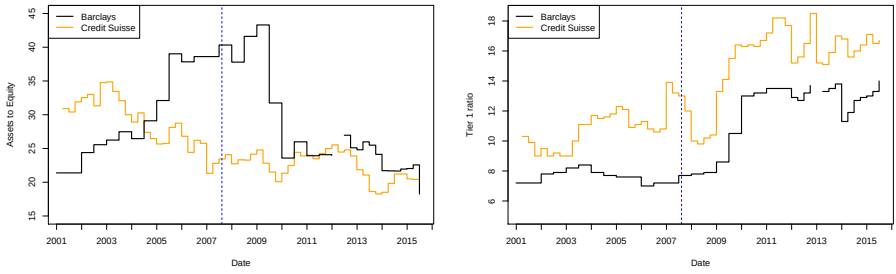
Figure 3.9: Indicator function for the event "P-value for the test using $T^{(A)}_{k_1,k_2,n_1,n_2}$ is below $\alpha/2 = 2.5\%$" for all possible choices of k_1 and k_2 for pre- and post-crisis negative log-returns for Barclays and Credit Suisse.

3.5 Relating statistical conclusions with economic indicators

Above we provided statistical indicators for a Black Swan event in financial return data linked with the recent financial crisis, measuring the probability for the experienced losses in view of the a priori return data, and by testing for significant differences in the scale parameters of the Pareto tail before and after the crisis. For Barclays the return period for the experienced loss as a result of the financial crisis is extremely large ánd we find a significant difference in the scale parameters before and after the crisis, and so we label Barclays as having experienced a Black Swan event during the recent crisis *in view of the pre-crisis return data only*. In contrast, for Credit Suisse the statistical significance is not met and the return period is more than 50 times smaller than in the Barclays case.

Of course one should be able to explain the vulnerability of a bank to such a financial crisis in terms of its economic parameters. At the time of the financial crisis, Barclays was a bank with an outspoken amount of leverage. Barclays' ratio of the assets to the equity base was almost twice as large compared to the leverage of Credit Suisse (Figure 3.10a). Credit Suisse had indeed much less assets for every dollar of equity. This made Credit Suisse less susceptible to a shock in the financial system.

When studying the Tier 1 ratio of both banks, the same conclusion holds (Figure 3.10b). This ratio relates the Tier 1 capital to the risk-weighted assets of a financial institution. Here, Barclays stands out again as a more vulnerable bank compared to Credit Suisse. Because of its vulnerability, Barclays witnessed a true Black Swan event, whereas this was not the case for Credit Suisse. This provides some explanation for the statistical conclusions obtained above.



(a) (b)

Figure 3.10: (a) Assets to equity ratio and (b) Tier 1 ratio (in %) for Barclays (black) and Credit Suisse (orange).

3.6 Conclusions

We investigated, based on EVT, if the recent financial crisis was a Black Swan event. More precisely, we were looking for a difference in tail behaviour before and after the crisis as indicated by the return periods for the experienced losses in view of the pre-crisis data, and tests for significant differences in the scale or shape parameters of the Pareto tail before and after the crisis. To test for a difference in scale, we developed new estimators for the scale parameter and provided asymptotic results for weakly-dependent data. The analysis indicated that Barclays can be considered as having experienced a Black Swan event whereas this is not the case for Credit Suisse. Economic indicators of both banks suggested that Barclays was indeed more vulnerable than Credit Suisse.

Another possible approach to investigate if the 2007-2008 financial crisis was a Black Swan event, is to test for change points as proposed by Dierckx and Teugels (2010). Their test generalises the likelihood approach of Csörgő and Horváth (1997) to an extreme value context. However, it can only be performed when the data are independent which is clearly not the case here. Dierckx and Teugels (2010) propose to overcome this problem by first declustering the data as described in Ferro and Segers (2003) and then performing their testing approach. Declustering the data will reduce the size of the dataset which results in lower testing power. This downside is not present when testing for Black Swans using our approach as all data can be used since dependence is taken into account in the test.

Chapter 4

Fitting tails affected by truncation

This chapter is based on

Beirlant, J., Fraga Alves, I. and Reynkens, T. (2017). Fitting Tails Affected by Truncation. *Electron. J. Stat.*, **11**(1), 2026–2065.

4.1 Introduction

Assessing the risk of rare events through estimation of extreme quantiles or corresponding return periods has been developed extensively. The previous chapter showed the use of extreme value theory (EVT) to model extreme events in finance. As mentioned in the previous chapter, the methodology on modelling the univariate upper tail of the distribution of such quantities is based on (3.1). For a random variable Y, condition (3.1) is equivalent to the convergence of the distribution of excesses (or peaks) over high thresholds t to the GPD: as t tends to the endpoint of the distribution of Y, then, with $\bar{F}$ the right tail function (RTF) or survival function of a given distribution,

$$P\left(\frac{Y-t}{\sigma_Y(t)} > y \,\middle|\, Y > t\right) = \frac{\bar{F}_Y(t+y\sigma_Y(t))}{\bar{F}_Y(t)} \to H_\xi(y) = -\ln G_\xi(y) = (1+\xi y)^{-1/\xi}, \tag{4.1}$$

where $\sigma_Y(t) > 0$, see also (3.3). Below we set $\sigma_Y(t) = \sigma_t$. Setting t at the $(k+1)$th largest observation $y_{n-k,n}$ for some $k \in \{1\ldots,n-1\}$ so that k data

points are larger than the threshold t, (4.1) leads to the estimator

$$\hat{p}_c = \frac{k}{n} H_{\hat{\xi}} \left(\frac{c - y_{n-k,n}}{\hat{\sigma}} \right) \tag{4.2}$$

of the tail probability $P(Y > c)$ for $c > 0$ large, where $(\hat{\xi}, \hat{\sigma})$ denote estimators for (ξ, σ_t). The modelling of extreme values and the estimation of tail parameters through the POT methodology has been discussed for instance in Embrechts et al. (1997), Coles (2001), Beirlant et al. (2004), and de Haan and Ferreira (2006).

Recently, Aban et al. (2006), Chakrabarty and Samorodnitsky (2012) and Beirlant et al. (2016a) have addressed the problem of using unbounded probability mass leading to levels that are unreasonably large or physically impossible. All of these papers consider cases with shape parameter $\xi > 0$. In Beirlant et al. (2016a) it was observed that the above mentioned extreme value methods based on the POT methodology, even when using a negative extreme value index, are not able to capture truncation at high levels. However, in several other fields, such as hydrology and earthquake magnitude modelling, the underlying distributions appear to be lighter tailed than Pareto. We propose an adaptation of the classical approach to truncated tails over the whole range of max-convergence (3.1) with $\xi > -0.5$ as in the original POT approach.

First, we revisit the diamond weight data considered in Verster et al. (2012). They note that the nature of metallurgical recovery processes in diamond mining may cause under recovery of large diamonds. Stones that are not recovered during this process end up at mine waste dumps. Mining companies want to investigate whether re-mining these dumps is profitable because of the recovery of large diamonds. They want to estimate the expected number of large diamonds above certain carat values c, and the original non-truncated values need to be reconstructed from the truncated data. In Figure 4.1, the Pareto QQ-plot of the available diamond weight data is presented. A curvature near the top data is visible which indicates that the distribution might indeed be truncated.

Second, we consider flows of the Molenbeek river in Erpe-Mere (Belgium). The data are peaks over threshold values from a full series of hourly flow measurements which was filtered to satisfy hydrological independence, see Willems (2009). Flooding can occur at high flow levels, and the measurements can thus be truncated. In Figure 4.2 the exponential QQ-plot is given, which exhibits a linear (i.e. exponential) pattern with again a curvature near the largest flows.

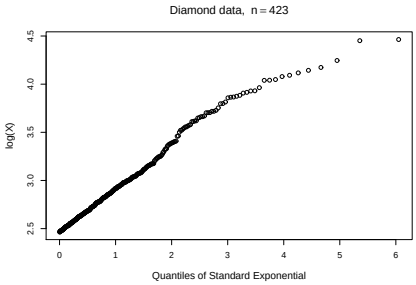


Figure 4.1: Pareto QQ-plot of diamond weight data from Verster et al. (2012).

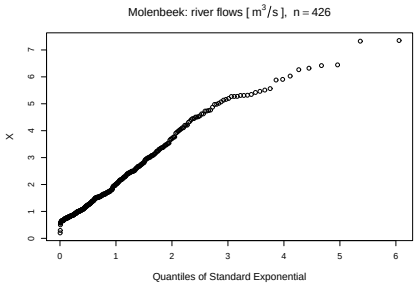


Figure 4.2: Exponential QQ-plot of the Molenbeek flow data.

Based on empirical evidence, it is often assumed that earthquake magnitudes follow the Gutenberg-Richter (GR) distribution (Gutenberg and Richter, 1956; Page, 1968) which is a doubly truncated exponential distribution. It has CDF

$$F(m) = \begin{cases} 0 & \text{if } m \leq t_M \\ \frac{(1-\exp(-\beta m))-(1-\exp(-\beta t_M))}{(1-\exp(-\beta T_M))-(1-\exp(-\beta t_M))} & \text{if } t_M < m < T_M \\ 1 & \text{if } m \geq T_M, \end{cases}$$

where $t_M > 0$ is the minimum possible magnitude, i.e. the lower truncation point, $T_M > t_M$ the maximum possible magnitude, i.e. the upper truncation point or endpoint, and $\beta > 0$ the rate parameter. Note that the Gutenberg-Richter distribution is not only derived empirically; e.g. Scholz (1968) discusses its relationship with earthquake physics. We hence expect that earthquake magnitudes are truncated with an underlying exponential distribution. In the next chapter we give a detailed analysis of earthquake magnitudes from Groningen.

We aim to provide a statistical model being able to approximate tail characteristics of distributions truncated at high levels. Moreover, the statistical estimation methods should also include the case of no-truncation in order for these methods to be useful and competitive both in cases with and without truncation. In the case of Pareto-type tails with $\xi > 0$ the proposed methods should also be compared with the methods which have been developed specifically for that sub-case. To this purpose we extend the classical POT technique with maximum likelihood estimation of the GPD parameters ξ and σ. Of course estimators for tail probabilities and extreme quantiles of a truncated distribution are to be discussed. Estimation of the endpoint T of a truncated distribution is of particular importance for earthquake magnitudes as we will illustrate in the

next chapter. Motivated by the diamond valuation example, we finally consider the problem of reconstructing quantiles of the underlying unobserved variable Y before truncation.

4.2 Model

Let Y denote a parent random variable with distribution function $F_Y(y) = P(Y \leq y)$, RTF $\bar{F}_Y(y) = 1 - F_Y(y)$, quantile function $Q_Y(p) = \inf\{y \mid F_Y(y) \geq p\}$ $(0 < p < 1)$, and tail quantile function $U_Y(v) = Q_Y(1 - \frac{1}{v})$ $(v > 1)$. We consider the upper truncated distribution from which independent and identically distributed data $X_1, X_2, \ldots, X_n$ are observed with, for some $T > 0$,

$$X =_d Y \mid Y < T. \tag{4.3}$$

The corresponding RTF is denoted by $\bar{F}_T(x) = P(X > x)$ and the tail quantile function is given by $U_T(u) = Q_T(1 - \frac{1}{u})$ $(u > 1)$. Then,

$$\bar{F}_T(x) = \frac{\bar{F}_Y(x) - \bar{F}_Y(T)}{1 - \bar{F}_Y(T)} = (1 + D_T)\bar{F}_Y(x) - D_T, \tag{4.4}$$

$$U_T(u) = U_Y\left(\frac{u}{F_Y(T)}\left[1 + uD_T\right]^{-1}\right) \tag{4.5}$$

$$= U_Y\left(\frac{1}{\bar{F}_Y(T)}\left[1 + \frac{1}{uD_T}\right]^{-1}\right), \tag{4.6}$$

where $D_T = \bar{F}_Y(T)/F_Y(T)$ equals the odds of the truncated probability mass under the untruncated distribution Y.

The goal is to provide a test for truncation and to estimate

- the model parameters ξ and $\sigma = \sigma_t$,
- the odds D_T,
- quantiles $Q_T(1 - p)$ (p small) of the truncated distribution and the truncation point $T = Q_T(1)$,
- tail probabilities $P(X > c)$ (c large) of the truncated distribution,
- and reconstruct quantile levels $Q_Y(1 - p)$ of the parent variable Y before truncation,

all on the basis of a pure random sample from X (possibly) truncated at some large T.

We assume that the distribution of Y satisfies (3.1) or, equivalently, (4.1). Condition (4.1) is also known to be equivalent to the following condition relating extreme quantile levels at $1-\frac{1}{vy}$ and $1-\frac{1}{y}$ close to the endpoint of the distribution: there exists a positive measurable function a such that

$$\lim_{y\to\infty} \frac{U_Y(vy)-U_Y(y)}{a(y)} = \frac{v^\xi - 1}{\xi}, \tag{4.7}$$

see (3.4). Here is $a(1/\bar{F}_Y(t_{k,n})) = \sigma_t$ where $t = t_{k,n} = U_T(n/k)$. The right hand side of (4.7) is to be read as $\ln v$ for $\xi = 0$. Moreover, for the specific case $\xi > 0$ of Pareto-type distributions, Y satisfies (3.5) and (3.6), and $\sigma_t \sim \xi t$ as $t \to \infty$. Furthermore, it is known that $\sigma_t/t \to 0$ when $\xi \le 0$.

Note that for a given T fixed, the tail of a truncated model X defined through (4.3) has an extreme value index $\xi_X = -1$, see for instance Figure 2.8 in Beirlant et al. (2004).

Truncation of a distribution Y satisfying (4.1) at a value T necessarily requires $t < T \to \infty$. The threshold t is mostly taken at the theoretical quantile $Q_T(1-\frac{k}{n}) = U_T(n/k)$, which in practice is estimated by the empirical quantile $X_{n-k,n}$. Given the fact that our model is only defined choosing $t = t_n, T = T_n \to \infty$ as the sample size $n \to \infty$, the underlying model depends on n and a triangular array formulation $X_{n1}, \ldots, X_{nn}$ of the observations should be used in order to emphasise the nature of the model. However, in statistical procedures as presented here, when a single sample is given, the notation $X_1, \ldots, X_n$ is more natural and will be used throughout.
The considered model is then given by

($\mathcal{M}$) For a sequence $T_n \to \infty$, $\{X_{n1}, \ldots, X_{nn}\} = \{X_1, \ldots, X_n\}$ are independent copies of a random variable $X = X_{T_n}$ where $X = X_{T_n}$ is distributed as $Y \,|\, Y < T_n$, with Y satisfying (4.1) or equivalently (4.7).

Now we consider the distribution of the POT values for the data of the truncated distribution under ($\mathcal{M}$):

$$P\left(\frac{X-t}{\sigma_t} > x \,\middle|\, X > t\right) = P\left(\frac{Y-t}{\sigma_t} > x \,\middle|\, t < Y < T\right) = \frac{P(Y > t + x\sigma_t) - P(Y > T)}{P(Y > t) - P(Y > T)}$$
$$= \frac{\frac{P(Y>t+x\sigma_t)}{P(Y>t)} - \frac{P(Y>T)}{P(Y>t)}}{1 - \frac{P(Y>T)}{P(Y>t)}}. \tag{4.8}$$

One can now consider two cases as $t, T \to \infty$:

- ($\mathcal{T}_t$) *Rough truncation with the threshold* $t = t_n$:

$$\frac{T-t}{\sigma_t} \to \kappa > 0, \tag{4.9}$$

and hence from (4.1) and with local uniform convergence in (4.1)

$$\frac{P(Y>T)}{P(Y>t)} \to (1+\xi\kappa)^{-1/\xi}. \tag{4.10}$$

This entails that for $x \in (0, \kappa)$

$$P\left(\frac{X-t}{\sigma_t} > x \,\middle|\, X > t\right) \to \frac{(1+\xi x)^{-1/\xi} - (1+\xi\kappa)^{-1/\xi}}{1-(1+\xi\kappa)^{-1/\xi}} =: \bar{F}_{\xi,\kappa}(x). \tag{4.11}$$

This corresponds to situations where the deviation from the Pareto behaviour due to truncation at a high value T will be visible in the data from t on, and the approximation of the POT distribution using the limit distribution in (4.11) appears more appropriate than with a simple GPD.

- ($\bar{\mathcal{T}}_t$) *Light truncation with the threshold* $t = t_n$*:* $\frac{P(Y>T)}{P(Y>t)} \to 0$.
This entails

$$P\left(\frac{X-t}{\sigma_t} > x \,\middle|\, X > t\right) \to (1+\xi x)^{-1/\xi}, \;\; 1+\xi x > 0. \tag{4.12}$$

Light truncation is introduced for mathematical completeness. But ($\bar{\mathcal{T}}_t$) means that the truncation is not really visible in the data above t, and the classical extreme value modelling without truncation is appropriate. Hence, it will be practically impossible to discriminate light truncation from no truncation (i.e. $T = \infty$).

Under ($\mathcal{T}_t$) with $t = t_{k,n} = U_T(n/k)$ we find from applying F_Y to both sides of (4.5) with $u = n/k$ that

$$\bar{F}_Y(t) = F_Y(T)\frac{1+(n/k)D_T}{n/k} = F_Y(T)\left(\frac{k}{n} + D_T\right),$$

from which, dividing by $\bar{F}_Y(T)$, we obtain

$$\frac{\bar{F}_Y(t)}{\bar{F}_Y(T)} = \frac{1}{D_T}\left(\frac{k}{n} + D_T\right),$$

while, using (4.1) and $(\mathcal{T}_t)$,

$$\frac{\bar{F}_Y(T)}{\bar{F}_Y(t)} \to (1+\xi\kappa)^{-1/\xi},$$

and hence under $(\mathcal{T}_t)$

$$\frac{k}{nD_T} \to (1+\xi\kappa)^{1/\xi} - 1. \tag{4.13}$$

Now in order to be able to construct extreme quantile estimators under $(\mathcal{T}_t)$, remark that from (4.7) with $vy = 1/p$, $y = 1/\bar{F}_Y(t)$ and $k_\xi(u) = (u^\xi - 1)/\xi$, we have as $t \to \infty$ and $\bar{F}_Y(t)/p \to C$ for some constant $C > 0$ that

$$\frac{Q_Y(1-p) - t}{\sigma_t} - k_\xi\left(\frac{\bar{F}_Y(t)}{p}\right) \to 0.$$

Hence, with (4.6) and $p = \bar{F}_Y(T)(1 + \frac{1}{uD_T})$ we obtain

$$\frac{U_T(u) - t}{\sigma_t} = \frac{U_Y\left(\frac{1}{\bar{F}_Y(t)}[1 + \frac{1}{uD_T}]^{-1}\right) - t}{\sigma_t} = k_\xi\left(\frac{\bar{F}_Y(t)}{\bar{F}_Y(T)[1 + \frac{1}{uD_T}]}\right) + o(1).$$

Using (4.13) and (4.1) with $y = \kappa$ we obtain under $(\mathcal{T}_t)$ that

$$\frac{\bar{F}_Y(t)}{\bar{F}_Y(T)} \sim (1+\xi\kappa)^{1/\xi} \sim 1 + \frac{k}{nD_T}.$$

Hence, we conclude that under $(\mathcal{T}_t)$ for $1/(uD_T) \to 0$

$$\frac{U_T(u) - t}{\sigma_t} - k_\xi\left(\frac{1 + \frac{k}{nD_T}}{1 + \frac{1}{uD_T}}\right) \to 0. \tag{4.14}$$

These derivations will motivate the proposed estimators of D_T and extreme quantiles $Q_T(1-p)$.

4.3 Inference

4.3.1 Estimators and goodness-of-fit

Estimation of the parameters (ξ, σ) in the classical POT without truncation is well-developed (Coles, 2001; Beirlant et al., 2004). Fitting the scaled GPD with RTF $\left(1 + \frac{\xi}{\sigma}x\right)^{-1/\xi}$ to the excesses $X - t$ given $X > t$ (based on (4.1))

using maximum likelihood is by far the most popular method in this respect. Here we rely on the generalisation (4.11) under $(\mathcal{T}_t)$, with t replaced by a random threshold $X_{n-k,n}$ and using the exceedances $E_{j,k} = X_{n-j+1,n} - X_{n-k,n}$ $(j = 1, 2, \ldots, k)$ for some $k \geq 2$. Substituting $E_{1,k}/\sigma$ for κ following (4.9), the log-likelihood is given by

$$\begin{aligned}
\ln L_{k,n}(\xi, \sigma) &= \ln \left(\prod_{j=2}^{k} \frac{\sigma^{-1} \left(1 + \frac{\xi}{\sigma} E_{j,k}\right)^{-(1/\xi)-1}}{1 - \left(1 + \frac{\xi}{\sigma} E_{1,k}\right)^{-1/\xi}} \right) \\
&= -(k-1) \ln \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{j=2}^{k} \ln \left(1 + \frac{\xi}{\sigma} E_{j,k}\right) \\
&\quad - (k-1) \ln \left(1 - \left(1 + \frac{\xi}{\sigma} E_{1,k}\right)^{-1/\xi}\right),
\end{aligned}$$

or, by reparametrising (ξ, σ) to (ξ, τ) with $\tau = \xi/\sigma$,

$$\begin{aligned}
\ln L_{k,n}(\xi, \tau) &= (k-1) \ln \tau - (k-1) \ln \xi - \left(1 + \frac{1}{\xi}\right) \sum_{j=2}^{k} \ln(1 + \tau E_{j,k}) \\
&\quad - (k-1) \ln \left(1 - (1 + \tau E_{1,k})^{-1/\xi}\right).
\end{aligned}$$

The partial derivatives are given by

$$\begin{aligned}
\frac{1}{k-1} \frac{\partial \ln L_{k,n}(\xi, \tau)}{\partial \xi} &= -\frac{1}{\xi} + \frac{1}{\xi^2} \frac{1}{k-1} \sum_{j=2}^{k} \ln(1 + \tau E_{j,k}) \\
&\quad + \frac{1}{\xi^2} \frac{(1 + \tau E_{1,k})^{-1/\xi} \ln(1 + \tau E_{1,k})}{1 - (1 + \tau E_{1,k})^{-1/\xi}}, \\
\frac{1}{k-1} \frac{\partial \ln L_{k,n}(\xi, \tau)}{\partial \tau} &= \frac{1}{\tau} - \left(1 + \frac{1}{\xi}\right) \frac{1}{k-1} \sum_{j=2}^{k} \frac{E_{j,k}}{1 + \tau E_{j,k}} \\
&\quad - \frac{1}{\xi} E_{1,k} \frac{(1 + \tau E_{1,k})^{-1-1/\xi}}{1 - (1 + \tau E_{1,k})^{-1/\xi}},
\end{aligned}$$

from which the likelihood equations defining the pseudo maximum likelihood estimators $(\hat{\xi}_k, \hat{\tau}_k)$ are obtained:

$$\frac{1}{k-1}\sum_{j=2}^{k}\ln(1+\hat{\tau}_k E_{j,k}) + \frac{(1+\hat{\tau}_k E_{1,k})^{-1/\hat{\xi}_k}\ln(1+\hat{\tau}_k E_{1,k})}{1-(1+\hat{\tau}_k E_{1,k})^{-1/\hat{\xi}_k}} = \hat{\xi}_k \tag{4.15}$$

$$\frac{1}{k-1}\sum_{j=2}^{k}\frac{1}{1+\hat{\tau}_k E_{j,k}} = \frac{1}{1+\hat{\xi}_k}\frac{1-(1+\hat{\tau}_k E_{1,k})^{-1-1/\hat{\xi}_k}}{1-(1+\hat{\tau}_k E_{1,k})^{-1/\hat{\xi}_k}}. \tag{4.16}$$

When computing $(\hat{\xi}_k, \hat{\tau}_k)$, one has to impose the model restrictions. In order to meet the restrictions $\sigma = \xi/\tau > 0$ and $1 + \tau E_{j,k} > 0$ for $j = 1, \ldots, k$, in our implementation we require the estimates of these quantities to be larger than the numerical tolerance value 10^{-10}.

An estimator of D_T now follows from taking $u = n$ in (4.14):

$$U_T(n) - U_T(n/k) \approx \sigma k_\xi \left(\frac{1+\frac{k}{nD_T}}{1+\frac{1}{nD_T}}\right).$$

Estimating $U_T(n) - U_T(n/k)$ by $E_{1,k}$ we obtain

$$\hat{D}_{T,k} = \max\left\{0, \frac{k}{n}\frac{(1+\hat{\tau}_k E_{1,k})^{-1/\hat{\xi}_k} - \frac{1}{k}}{1-(1+\hat{\tau}_k E_{1,k})^{-1/\hat{\xi}_k}}\right\}. \tag{4.17}$$

Similarly taking $u = 1/p$ in (4.14) with $np/k \to 0$, we obtain an estimator for $Q_T(1-p)$:

$$\hat{Q}_{T,k}(1-p) = X_{n-k,n} + \frac{1}{\hat{\tau}_k}\left(\left[\frac{\hat{D}_{T,k} + \frac{k}{n}}{\hat{D}_{T,k} + p}\right]^{\hat{\xi}_k} - 1\right). \tag{4.18}$$

Based on (4.2) and (4.4) an estimator for tail probabilities $P(X > c)$ can be derived:

$$\hat{p}_{T,k}(c) = (1+\hat{D}_{T,k})\frac{k}{n}\left(1+\hat{\tau}_k(c - X_{n-k,n})\right)^{-1/\hat{\xi}_k} - \hat{D}_{T,k}. \tag{4.19}$$

Note that all proposed estimators from (4.15), (4.16), (4.18) and (4.19) are direct generalisations of the classical POT estimators under no-truncation which are obtained by setting $\hat{D}_{T,k}$ equal to 0.

From (4.5) it follows that when $p-(1-p)D_T > 0$, or $p > D_T/(1+D_T) = \bar{F}_Y(T)$

$$Q_Y(1-p) = Q_T\left((1-p)(1+D_T)\right) = Q_T\left((1-(p-(1-p)D_T)\right),$$

from which the following estimator reconstructing $Q_Y(1-p)$ of the parent distribution Y emerges:

$$\hat{Q}_{Y,k}(1-p) = \hat{Q}_{T,k}\left(1-[p-(1-p)\hat{D}_{T,k}]\right)$$
$$= X_{n-k,n} + \frac{1}{\hat{\tau}_k}\left(\left[\frac{\hat{D}_{T,k}+\frac{k}{n}}{p(\hat{D}_{T,k}+1)}\right]^{\hat{\xi}_k} - 1\right). \tag{4.20}$$

In the specific case $\xi > 0$ the estimators developed above can be compared with those developed in Beirlant et al. (2016a) for this special Pareto-type case:

$$H_{k,n} = \hat{\xi}_k^+ + \frac{R_{k,n}^{1/\hat{\xi}_k^+}\ln R_{k,n}}{1-R_{k,n}^{1/\hat{\xi}_k^+}}, \tag{4.21}$$

$$\hat{D}_{T,k}^+ = \max\left\{0, \frac{k}{n}\frac{R_{k,n}^{1/\hat{\xi}_k^+} - \frac{1}{k}}{1-R_{k,n}^{1/\hat{\xi}_k^+}}\right\}, \tag{4.22}$$

$$\ln \hat{Q}_{T,k}^+(1-p) = \ln X_{n-k,n} + \hat{\xi}_k^+ \ln\left(\frac{\hat{D}_{T,k}^+ + \frac{k}{n}}{\hat{D}_{T,k}^+ + p}\right), \tag{4.23}$$

with $H_{k,n} = \frac{1}{k}\sum_{j=1}^{k}\ln X_{n-j+1,n} - \ln X_{n-k,n}$ the Hill (1975) statistic, and $R_{k,n} = X_{n-k,n}/X_{n,n}$.

Of course, in practice there is a clear need for detecting rough truncation. Let $(\bar{\mathcal{T}}_k)$ and $(\mathcal{T}_k)$ denote light and rough truncation with the thresholds $X_{n-k,n}$. A test for

$$H_{0,k} : (\bar{\mathcal{T}}_k) \text{ versus } H_{1,k} : (\mathcal{T}_k)$$

can be constructed generalising the goodness-of-fit test which was proposed by Aban et al. (2006) within a Pareto context, rejecting $H_{0,k}$ at asymptotic level $q \in (0,1)$ when

$$T_{k,n} := k\left(1+\hat{\tau}_k E_{1,k}\right)^{-1/\hat{\xi}_k} > \ln(1/q), \tag{4.24}$$

while the P-value is given by $e^{-T_{k,n}}$, as under $H_{0,k}$, $T_{k,n}$ approximately follows a standard exponential distribution as will be shown in Theorem 4.3 below.

4.3.2 Simulation study

The authors have performed an extensive simulation study concerning all the proposed estimators for different distributions of Y. We compare the results

with the results from a Pareto analysis $\hat{\xi}_k^+$ and $\hat{Q}_{T,k}^+(1-p)$ (Aban et al., 2006; Beirlant et al., 2016a), with the classical POT maximum likelihood results denoted by $\hat{\xi}_k^\infty$, $\hat{Q}_k^\infty(1-p)$, and with the classical moment estimators (Dekkers et al., 1989)

$$\hat{\xi}_k^{\text{Mom}} = M_k^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{\left(M_k^{(1)}\right)^2}{M_k^{(2)}}\right)^{-1}, \tag{4.25}$$

$$\hat{Q}_k^{\text{Mom}}(1-p) = X_{n-k,n} + X_{n-k,n} M_k^{(1)} \left(1 - \hat{\xi}_k^{\text{Mom}}\right) \frac{\left(\frac{k}{np}\right)^{\hat{\xi}_k^{\text{Mom}}} - 1}{\hat{\xi}_k^{\text{Mom}}}, \tag{4.26}$$

with $M_k^{(j)} = \frac{1}{k}\sum_{l=1}^{k} \ln^j \left(X_{n-l+1,n}/X_{n-k,n}\right)$, $j = 1, 2$. In Appendix B.2 we give a selection from these simulation results for Y following the standard Pareto distribution, the standard lognormal distribution, the standard exponential distribution, and the GPD with RTF $H_{-0.2}$. For each setting, 1000 samples for X of size 500 were generated where we consider different levels of truncation: $T = Q_Y(0.975)$, $T = Q_Y(0.99)$ and $T = Q_Y(1)$. Note that the last case corresponds to no truncation, or $X =_d Y$. The samples were generated using inverse transform sampling with the quantile function $Q_T(p) = Q_Y(pF_Y(T))$ (which can easily be deduced from (4.4)).

To show the performance of the test for truncation, we plot the average P-values over the 1000 simulations as a function of k in the first columns of Figures B.1–B.4 (full line). Additionally, the median (dashed line), first quartile (dotted line) and third quartile (dotted line) of the P-values over the 1000 simulations are also plotted as a function of k. This corresponds to the box of the boxplot of P-values as a function of k. Finally, we add horizontal lines (dash-dotted line) indicating the standard significance levels of 1% and 5%. When truncation is present ($T = Q_Y(0.975)$ or $T = Q_Y(0.99)$), the average P-values show that the test rejects the null hypothesis of no truncation when k is large enough. For the standard exponential, standard lognormal and GPD(-0.2,1) truncated at $T = Q_Y(0.99)$, the average P-value is higher than, or just below, the 5% significance level, even for high values of k. However, when looking at the median values and the third quartile, we see that the majority, and sometimes more than 75%, of the P-values are below the 5% significance level. When the data are not truncated, i.e. $X =_d Y$, the P-values are on average always well above the considered significance levels, hence correctly not rejecting the null hypothesis. The first quartile of the P-values is also above the 5% significance level, except for smaller values of k. Note that when we look at $Y \sim \text{GPD}(-0.2, 1)$, Y itself is bounded by $-\sigma/\xi = 5$, but still $X =_d Y$ when we set $T = Q_Y(1)$. The simulation results show that the test performs as expected: rejecting the null

hypothesis when $T = Q_Y(0.975)$ or $T = Q_Y(0.99)$, and not rejecting the null hypothesis when $T = Q_Y(1)$.

Concerning the estimation of ξ, see the second and third columns in Figures B.1–B.4, the behaviour of $\hat{\xi}_k$ in the standard Pareto case exhibits a slightly smaller bias but quite a larger variance compared to $\hat{\xi}^+_{T,k}$ from Aban et al. (2006) and Beirlant et al. (2016a) which was constructed exclusively for the case $\xi > 0$. The classical POT and moment estimators exhibit large bias under truncation, as they tend to -1 when the threshold tends to $x_{n,n}$. The mean squared error of $\hat{\xi}_k$ is comparable to the mean squared error (MSE) of these estimators for $k \geq 200$. In case of no truncation the bias of $\hat{\xi}_k$ is the smallest for $k \geq 100$ while the mean squared error is the worst of the four estimators. When $\xi \leq 0$, the estimator $\hat{\xi}^+_{T,k}$ from the Pareto analysis is breaking down as can be expected whereas the difference between the classical estimators and the newly proposed POT estimator is small for $k \geq 200$ in case $\xi = 0$ and $k \geq 300$ in the case $\xi < 0$. In all presented cases, $\hat{\xi}_k$ compares well for k sufficiently large with the classical estimators when there is no truncation. Note that all estimators have a large bias for the (truncated) lognormal distribution. As can clearly be seen, the bias of all estimators decreases as truncation becomes lighter, or when there is no truncation, as expected. Moreover, the stable area of the $\hat{\xi}_k$ estimates starts for smaller values of k when the truncation point gets larger.

Concerning the estimation of $Q_T(1-p)$, see Figures B.5–B.12 with $p = 0.01$ and 0.005 and $T = Q_Y(0.975), Q_Y(0.99)$, the estimator $\hat{Q}_{T,k}(1-p)$ has the smallest bias, uniformly over all distributions and values of p considered, while the MSE values are always comparable with the best performing estimators. Even in case of no truncation $\hat{Q}_{T,k}(1-p)$ does not lose too much accuracy in comparison with the classical maximum likelihood (ML) estimator.

4.3.3 Asymptotic results

Here we present the asymptotic normality of $(\hat{\xi}, \hat{\tau})$ and $\hat{Q}_{T,k}(1-p)$ under rough truncation, and the asymptotic null distribution of the goodness-of-fit test statistic $T_{k,n}$. The proofs are provided in Appendix B.1.

We assume a second-order remainder relation in (4.7) as in Theorem 3.4.2 in de Haan and Ferreira (2006): with $\xi > -\frac{1}{2}$,

$$\lim_{t\to\infty} \frac{\frac{U_Y(tx)-U_Y(t)}{a_Y(t)} - \frac{x^\xi - 1}{\xi}}{A(t)} = \Psi_{\xi,\rho}(x) \text{ for all } x > 0, \tag{4.27}$$

where

$$\Psi_{\xi,\rho}(x) = \int_1^x s^{\xi-1} \int_1^s u^{\rho-1}\, du\, ds,$$

with $\rho \leq 0$. Furthermore, we introduce the notations $b_{T,k,n} := \frac{k+1}{(n+1)D_T}$, $a_{T,k,n} := a_Y\left(1/(\bar{F}_Y(T)(1+b_{T,k,n}))\right)$, and we denote the limit of $k/(nD_T)$ under rough truncation as derived in (4.13) by $\beta := (1+\xi\kappa)^{1/\xi} - 1$.

Theorem 4.1. *Let $X_1, X_2, \ldots,$ be i.i.d. random variables with distribution function F_T following (4.4) where U_Y satisfies (4.27). Let $n, k = k_n \to \infty$, $\frac{k}{n} \to 0$, $T \to \infty$. Then, under $(\mathcal{T}_t)$ we have that as $\sqrt{k}A(1/[\bar{F}_Y(T)(1+b_{T,k,n})]) \to \lambda$, with λ finite,*

$$\sqrt{k}\left(\hat{\xi}_k - \xi, \hat{\tau}_k a_{T,k,n} - \xi\right)' = \mathcal{I}_\beta^{-1}\mathbf{N}_{\xi,\beta} + \lambda\mathcal{I}_\beta^{-1}\mathbf{f}_{\xi,\beta,\rho} + o_p(1)\mathbf{1},$$

where

$$\mathcal{I}_\beta = \begin{pmatrix} 1 - \frac{1+\beta}{\beta^2}\ln^2(1+\beta) & \frac{1}{\xi}\left[-\frac{\xi}{1+\xi}\frac{1+\beta}{\beta}(1-(1+\beta)^{-1-\xi}) + \frac{1+\beta}{\beta^2}\ln(1+\beta)(1-(1+\beta)^{-\xi})\right] \\ -\frac{1}{\xi}\left[-\frac{\xi}{1+\xi}\frac{1+\beta}{\beta}(1-(1+\beta)^{-1-\xi}) + \frac{1+\beta}{\beta^2}\ln(1+\beta)(1-(1+\beta)^{-\xi})\right] & -\frac{1}{\xi\beta}\left[\frac{\xi}{1+2\xi}(1+\beta)(1-(1+\beta)^{-1-2\xi}) - \frac{1+\beta}{\beta}\frac{1}{\xi}(1-(1+\beta)^{-\xi})^2\right] \end{pmatrix},$$

$$\mathbf{N}_{\xi,\beta} = \frac{\beta}{1+\beta}\begin{pmatrix} \xi\int_0^1 W_n(u)\left(\frac{1+u\beta}{1+\beta}\right)^{-1} du \\ -\xi W_n(1)\left(-\frac{(1+\beta)^{1-\xi}\ln(1+\beta)}{\beta^2} + \frac{\xi(1+\beta)^{-\xi}+(1+\beta)}{(1+\xi)\beta}\right) \\ \xi(1+\xi)\int_0^1 W_n(u)\left(\frac{1+u\beta}{1+\beta}\right)^{-1+\xi} du \\ -W_n(1)\left(\frac{\xi(1+\xi)(1+\beta)}{(1+2\xi)\beta}(1-(1+\beta)^{-1-2\xi}) - \frac{(1+\beta)^{1-\xi}}{\beta^2}(1-(1+\beta)^{-\xi})\right) \end{pmatrix},$$

and

$$\mathbf{f}_{\xi,\beta,\rho} = \begin{pmatrix} \xi\int_0^1 \Psi_{\xi,\rho}(\frac{1+\beta}{1+u\beta})\left(\frac{1+u\beta}{1+\beta}\right)^{\xi} du \\ -\xi\Psi_{\xi,\rho}(1+\beta)(1+\beta)^{-\xi}\left(\frac{(1+\beta)\ln(1+\beta)}{\beta^2} - \frac{1}{\beta}\right) \\ \xi(1+\xi)\int_0^1 \Psi_{\xi,\rho}(\frac{1+\beta}{1+u\beta})\left(\frac{1+u\beta}{1+\beta}\right)^{2\xi} du \\ -\Psi_{\xi,\rho}(1+\beta)\frac{(1+\beta)^{1-\xi}}{\beta^2}(1-(1+\beta)^{-\xi}) \end{pmatrix},$$

for a sequence of Brownian motions $\{W_n(s) \,|\, s \geq 0\}$.

Under $(\bar{\mathcal{T}}_t)$ the asymptotic result for $(\hat{\xi}_k, \hat{\tau}_k)$ can be checked to be identical to that of the classical ML estimators under no truncation as given in Theorem 3.4.2 in de Haan and Ferreira (2006). Note that the information matrix $\mathcal{I}_\beta$ equals 0 when $\kappa = 0$, or equivalently $\beta = 0$, so that the asymptotic variances are unbounded in such case. In practice this induces large variances for smaller values of k. This also appears in Figures B.1–B.4. Fortunately, the bias stays reasonably small for larger values of k, as can be deduced for instance in case of the lognormal distribution.

In order to state the asymptotic result for the quantile estimator $\hat{Q}_{T,k}(1-p)$ with $p = p_n \to 0$, we use the notation $d_n = k/(np_n)$. Furthermore, we will use the result that when U_Y satisfies (4.27), we have that

$$\lim_{t\to\infty} \frac{\frac{a_Y(tx)}{a_Y(t)} - x^\xi}{A(t)} = Cx^\xi \frac{x^\rho - 1}{\rho} \tag{4.28}$$

for some constant C (see B.3.4 in de Haan and Ferreira (2006)).

Theorem 4.2. *Let $X_1, X_2, \ldots,$ be i.i.d. random variables with distribution function F_T following (4.4) where U_Y satisfies (4.27). Let $n, k = k_n \to \infty$, $\frac{k}{n} \to 0$, $T \to \infty$, $p = p_n \to 0$ and $np_n/\sqrt{k} \to 0$. Then, under $(\mathcal{T}_t)$ we have that*

$$\begin{aligned}
&\frac{\left(\hat{Q}_{T,k}(1-p) - Q_T(1-p)\right)}{a_Y\left(\frac{1}{\bar{F}_Y(T)}\right)} \\
&\quad = -\frac{\beta}{k}(E-1) + O_p\left(\frac{1}{k^2} \vee \frac{1}{d_n^2}\right) \\
&\qquad - \beta\left(\frac{1}{d_n} - \frac{1}{k}\right)\left[A\left(\frac{1}{\bar{F}_Y(T)}\right) C \frac{(1+\beta)^{-\rho} - 1}{\rho}\right. \\
&\qquad\qquad + \left(\frac{\hat{\xi}_k}{\hat{\tau}_k}\frac{1}{a_{T,k,n}} - 1\right) \\
&\qquad\qquad - \left(\hat{\xi}_k - \xi\right)\frac{1}{\xi}\frac{(1+\beta)\ln(1+\beta)}{\beta} \\
&\qquad\qquad \left. + (\hat{\tau}_k a_{T,k,n} - \xi)\frac{1-(1+\beta)^{-\xi}}{\xi}\left(1 + \frac{1+\beta}{\xi\beta}\right)\right.
\end{aligned}$$

$$+ (1+\beta)^{-\xi}\left(\frac{1+\beta}{\beta}+\xi\right)$$
$$\times\left(-\frac{W_n(1)}{\sqrt{k}}+A\left(\frac{1}{\bar{F}_Y(T)}\right)(1+\beta)^{-\xi}\Psi_{\xi,\rho}(1+\beta)\right)\Bigg],$$

where E is a standard exponential random variable and $\{W_n(s)\,|\,s\geq 0\}$ a sequence of Brownian motions.

This result should be compared to Theorem 4.3.1 in de Haan and Ferreira (2006) stating the basic asymptotic result for the quantile estimator based on the classical ML estimators under no truncation. Note that under $(\mathcal{T}_t)$ the rate of the stochastic part in the asymptotic representation is $O_p(1/k)$ rather than the classical $O_p(1/\sqrt{k})$.

Theorem 4.3. *Let $X_1, X_2, \ldots,$ be i.i.d. random variables with distribution function F_T following (4.4) where U_Y satisfies (4.27). Let $n, k = k_n \to \infty$, $\frac{k}{n}\to 0$, $T\to\infty$. Then, under $(\bar{\mathcal{T}}_k)$ with $nD_T\to 0$ we have that*

$$T_{k,n} =_d E(1+o_p(1))$$

where E is a standard exponential random variable.

4.4 Case studies

Concerning the diamond data introduced in Figure 4.1, $\hat{\xi}_k$ and $\hat{\xi}_k^+$, respectively $\hat{D}_{T,k}$ and $\hat{D}_{T,k}^+$, correspond well for $k\geq 250$ and lead to a Pareto fit with extreme value index around 0.5 and a truncation odds D_T around 0.02. The goodness-of-fit test rejects light truncation for $k\geq 110$. Reconstructing $Q_Y(0.99)$, the 99% quantile of the original non-truncated diamond weights, with $\hat{Q}_{Y,k}(0.99)$ and $\hat{Q}_{Y,k}^+(0.99)$ leads to a value of 120 cts at $k=250$.

Finally, with the Molenbeek data, the goodness-of-fit test and the fit of the proposed truncation model on the exponential QQ-plot on the top 100 data, indicate that this Y belongs to the Gumbel domain with an odds D_T around 0.02. Here, the Pareto domain estimators $\hat{\xi}_k^+$ and $\hat{D}_{T,k}^+$ clearly do not show a stable pattern as a function of k. Estimation of $Q_T(0.99)$ leads to a value $\hat{Q}_{T,100}(0.99)=6.75\ m^3/s$.

4.5 Conclusions

We proposed a general tail estimation approach for cases where truncation affects the ultimate right tail of the distribution. We motivated the importance of this problem using applications from hydrology and geology. The proposed estimators of the extreme value index, and quantiles of the truncated and underlying non-truncated distribution, in most cases compare well with the best performing alternatives, even in case there is no truncation. The proposed estimator of extreme quantiles of a truncated distribution is performing uniformly best. While the alternative procedures sometimes break down in at least one situation, our proposals remain always useful for large enough k. Hence, in addition to the existing methods, this method can be an interesting extra tool when analysing tails.

In the next chapter, we will use the proposed methodology to estimate the maximum possible earthquake magnitude in Groningen.

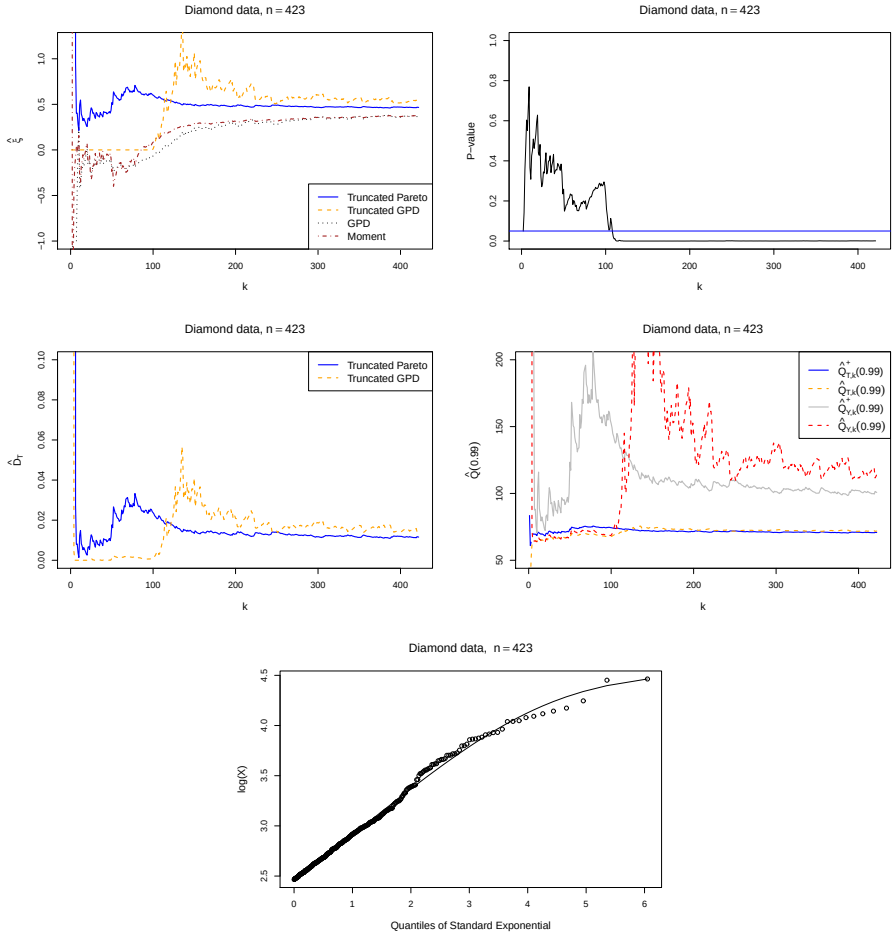


Figure 4.3: Diamond data: $\hat{\xi}_k^+$, $\hat{\xi}_k$, $\hat{\xi}_k^\infty$ and $\hat{\xi}_k^{\mathrm{Mom}}$ (top left); P-values for test for truncation (top right); $\hat{D}_{T,k}^+$ and $\hat{D}_{T,k}$ (middle left); $\hat{Q}_{T,k}^+(0.99)$, $\hat{Q}_{T,k}(0.99)$, $\hat{Q}_{Y,k}^+(0.99)$ and $\hat{Q}_{Y,k}(0.99)$ (middle right); Pareto QQ-plot with fit based on $k = 250$ largest weights (bottom).

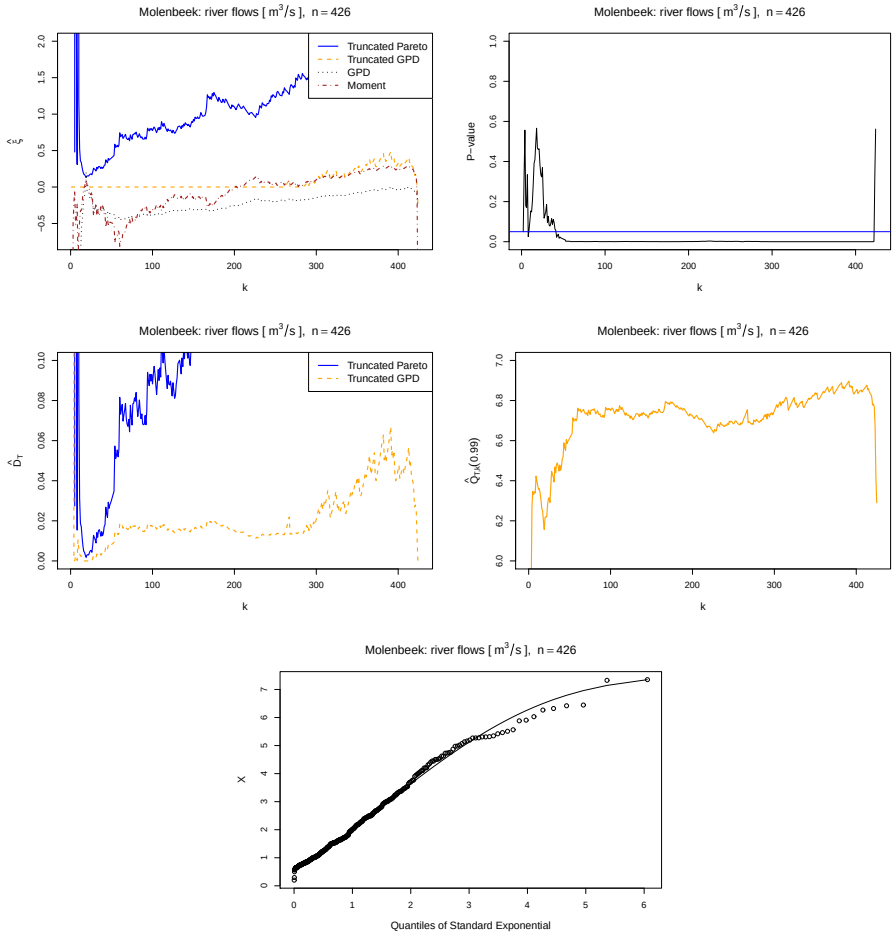


Figure 4.4: Molenbeek flow data: $\hat{\xi}_k^+$, $\hat{\xi}_k$, $\hat{\xi}_k^\infty$ and $\hat{\xi}_k^{\mathrm{Mom}}$ (top left); P-values for test for truncation (top right); $\hat{D}_{T,k}^+$ and $\hat{D}_{T,k}$ (middle left); $\hat{Q}_{T,k}(0.99)$ (middle right); exponential QQ-plot with fit based on $k = 100$ largest flows (bottom).

Chapter 5

Estimating the maximum possible earthquake magnitude in Groningen

5.1 Introduction

Under the province of Groningen lies one of the largest gas fields in the world. The reservoir lies at a depth of 3 km in Rotliegend sandstone and contains an estimated 2800 billion cubic metres of gas. Since production started in 1963, around 2000 billion cubic metres of gas has been produced up to 2012 by the NAM *(Nederlandse Aardolie Maatschappij)*, a partnership between Shell and ExxonMobil. As a result of its participation in NAM and taxes, the Dutch government typically receives 70% of the profit from the Groningen gas field, although in some periods this can be even as high as 90% (van der Voort and Vanclay, 2015).

Despite the economic advantages of the gas extraction on the Dutch government finances, there is also a serious drawback. Since 1986, the gas extraction induced earthquakes in the, otherwise mostly aseismic, northern part of the Netherlands, and especially in the province of Groningen. When the gas is extracted, the porous layer of sandstone, in which it is contained, compacts. Normally, this happens gradually and the surface subsides without causing problems. However, when this process happens close to fault lines, the sandstone layers can locally compact differently which causes earthquakes (van Eck et al., 2006; van der

Voort and Vanclay, 2015). As a consequence of these earthquakes, houses have been damaged, and the NAM has paid around 200 million euro of compensation up to 2014. Moreover, several thousands of houses need to be reinforced to avoid serious damage in a future earthquake. van Eck et al. (2006) also mention other social impacts of the earthquake including declining house prices, and concerns about breaching of the dykes in the gas field area in case of a large earthquake.

An important tool when investigating damage caused by earthquakes, is the magnitude of the earthquake. It is directly connected to the seismic energy of the earthquake at the epicentre (see (5.4) below). We look here at magnitudes expressed on the Richter scale. The maximal observed magnitude in Groningen is 3.6 which occurred on 16 August 2012 near the village of Huizinge (municipality of Loppersum).

Connected to the magnitudes is the intensity of the quake which depends on the location where it is measured: at the surface above the epicentre, etc. Since the earthquakes in Groningen occur in shallow sandstone layers, around 3 km depth, an earthquake with a relatively small magnitude can still have a high intensity. Maximal intensities corresponding to a shallow earthquake with magnitude between 4 and 5 will probably be in the range VI to VII on the European macroseismic scale (EMS-98) (Dost and Kraaijpoel, 2013). This corresponds to (European Seismological Commission, 1998):

- **Slightly damaging (VI)**: Objects on walls fall. Slight damage to buildings. Fine cracks in plaster and small pieces of plaster fall.
- **Damaging (VII)**: Furniture is shifted and many objects fall from shelves. Many buildings suffer slight to moderate damage. Cracks in walls, partial collapse of chimneys.

For the 2012 Huizinge earthquake, intensity VI was measured less than 4 km from the epicentre (Dost and Kraaijpoel, 2013).

The, area-characteristic, maximum possible earthquake magnitude T_M is required by the earthquake engineering community, disaster management agencies and the insurance industry. This is the maximum magnitude of an earthquake that can be generated by the geological structure of the area (Sintubin, 2016). This quantity thus only depends on the tectonic properties of the area, i.e. type of soil, faults, etc., and not on the evolution of the induced seismic activity. For the Groningen area, this means that it is independent of the production regime of the gas field. In contrast, the maximum expected earthquake magnitude during a certain time period does not only depend on

the tectonic properties, and hence on T_M, but also on the production regime of the gas field during that time period.

Based on magnitude data, we try to estimate the maximum possible earthquake magnitude. Several estimates for the Groningen area have been made by the KNMI *(Koninklijk Nederlands Meteorologisch Instituut)*: 3.3 in 1995, 3.8 in 1998 and 3.9 in 2004 (see e.g. Zöller and Holschneider, 2016b). In April 2016, a workshop was held in Amsterdam to provide an estimate for the maximum possible earthquake magnitude in Groningen, see NAM (2016) for an overview of the results. The range of maximum magnitude estimates for *induced* earthquakes provided by the experts is 3.8 to 5 (van den Beukel, 2016). The fault movements of these earthquakes are contained in the gas field, or propagate limited, i.e. less than 500 m, outside the gas field. If the fault movements of induced earthquakes propagate further outside of the gas field, they are called *triggered* earthquakes. These earthquakes can have larger magnitudes since they release tectonic tension that is built up in existing faults outside the gas field. For the moment, no scientific evidence has been found by geophysical experts that triggered earthquakes can occur in Groningen (Sintubin, 2016; van den Beukel, 2016). However, if they would occur, the experts estimate that the maximum magnitude can be as high as 7.25 (van den Beukel, 2016).

The estimation of T_M is of course an extreme value problem. Using the techniques that are studied in the previous chapter, we can provide an estimate for the maximum possible earthquake magnitude in Groningen. We compare this with two other EVT-based estimators: the estimator of Beirlant et al. (2016a) and the estimator of Fraga Alves et al. (2017). Other EVT-based estimators using the moment estimator (Dekkers et al., 1989) or the POT approach have also been proposed (see e.g. Beirlant et al., 2004; de Haan and Ferreira, 2006). Einmahl and Magnus (2008) use these techniques to estimate endpoints in records in athletics. As could be seen in the simulations in the previous chapter, these estimators for high quantiles, and thus also for the endpoint, perform worse for truncated distributions than the approach of the previous chapter and the estimator of Beirlant et al. (2016a). Therefore, the moment and POT endpoint estimators are omitted in this chapter. As suggested by Zöller and Holschneider (2016a), we also look at upper confidence bounds for the endpoint to give an idea about the uncertainty for the endpoint estimates. For this purpose, we use the asymptotic techniques of the previous chapter and Beirlant et al. (2016a), and the results from Fraga Alves et al. (2017).

The EVT-based estimators for the endpoint have received no attention yet in the geophysical literature, where several parametric and non-parametric estimators for the endpoint can be found. We give an overview of several non-parametric endpoint estimators as discussed in Kijko and Singh (2011). As mentioned in the previous chapter, it is often assumed that earthquake magnitudes follow

the GR distribution. Several parametric estimators have been proposed based on this distribution, see e.g. Pisarenko et al. (1996) and Raschke (2012). In this chapter, we only look at the parametric Kijko-Sellevol estimator (Kijko and Sellevoll, 1989). Moreover, we also look at a parametric upper bound for the maximum earthquake magnitude based on the GR distribution (Holschneider et al., 2011). Zöller and Holschneider (2016b) applied this technique to data from Groningen. Note that Bayesian estimators for the maximum earthquake magnitude have also been considered, see e.g. Cornell (1994), Holschneider et al. (2011) and Kijko (2012).

Zöller and Holschneider (2016b) also provide estimates for the maximum expected earthquake magnitude, for different production regimes, using Bayesian methods (Holschneider and Zöller, 2014). It is important to note that we do not try to estimate this quantity, but only look at estimates for the time-independent maximum possible earthquake magnitude.

In the next section, we discuss the different endpoint estimators that can be used to estimate the maximum possible earthquake magnitude. In Section 5.3, we apply these methods to estimate the maximum possible earthquake magnitude in Groningen. Moreover, we also discuss upper confidence bounds for this quantity. Afterwards, we compare the performance of the EVT-based estimators with those of the geophysical literature using simulations from the GR distribution.

5.2 Overview of estimators

We now discuss the different types of endpoint estimators: based on EVT in Section 5.2.1, non-parametric estimators as discussed in Kijko and Singh (2011) in Section 5.2.2 and the parametric Kijko-Sellevol estimator in Section 5.2.3. We only give limited details for the estimators from the geophysical literature as they are not in the main scope of this thesis. More details can be found in Kijko and Singh (2011), and for each estimator we will refer to the corresponding equation in this paper.

5.2.1 EVT-based estimators

We consider three EVT-based estimators of the endpoint: the truncated GPD estimator using the framework from the previous chapter, the truncated Pareto estimator of Beirlant et al. (2016a) and the FAN estimator of Fraga Alves et al. (2017).

Denote the ordered sample of magnitudes as $M_{1,n} \leq \ldots \leq M_{n,n}$. An overview of the used notation regarding the EVI and the endpoint can be found in Table 5.1.

Variable	EVI	Endpoint	Parent variable	EVI of parent variable
Magnitude M	ξ_M	T_M	Y with $M =_d Y \,\vert\, Y < T_M$	ξ
Energy E	ξ_E	T_E	Y_E with $E =_d Y_E \,\vert\, Y_E < T_E$	ξ_{Y_E}

Table 5.1: Magnitude and energy: overview of notation.

Truncated GPD estimator

We can estimate the endpoint of the magnitudes using the techniques from the previous chapter. Setting $p = 0$ in (4.18) one obtains an estimator for the truncation point T_M:

$$\hat{T}_k^M = M_{n-k,n} + \frac{1}{\hat{\tau}_k}\left[\left(\frac{1-\frac{1}{k}}{(1+\hat{\tau}_k(M_{n,n}-M_{n-k,n}))^{-1/\hat{\xi}_k}-\frac{1}{k}}\right)^{\hat{\xi}_k} - 1\right], \quad (5.1)$$

with $\hat{\xi}_k$ and $\hat{\tau}_k$ the estimates for ξ and τ obtained from (4.15) and (4.16). We denote this estimator by *Truncated GPD*. Note that, as in the previous chapter, ξ is the EVI of Y, the parent variable of M, see Table 5.1.

Using Theorem 4.2 with $p = 0$, we obtain an approximate $100(1-\alpha)\%$ upper confidence bound for T_M:

$$\hat{T}_k^M - (\ln\alpha + 1)\frac{\frac{k+1}{(n+1)\hat{D}_{T,k}}}{k+1}\hat{a}_Y\left(\frac{1}{\bar{F}_Y(T)}\right) \quad (5.2)$$

where second order terms have been omitted. Here, $\hat{a}_Y\left(\frac{1}{\bar{F}_Y(T)}\right) = \left(1+\frac{k+1}{(n+1)\hat{D}_{T,k}}\right)^{\hat{\xi}_k}\frac{\hat{\xi}_k}{\hat{\tau}_k}$, and $\hat{D}_{T,k}$ is the estimate for D_T, see (4.17).

Truncated Pareto estimator

The endpoint estimator of Beirlant et al. (2016a) is only suitable for truncated Pareto-type tails. Since we expect, based on the GR distribution, that the earthquake magnitudes have a truncated exponential-like distribution, we cannot apply this estimator to the magnitudes directly. Instead we use following

relationship between the earthquake magnitude (expressed on the Richter scale) and the energy from seismic waves (expressed in MJ):

$$E = 2 \times 10^{1.5(M-1)} = \exp(\ln 2 + (M-1)1.5 \ln 10), \tag{5.3}$$

or reversely

$$M = \frac{\log_{10}\left(\frac{E}{2}\right)}{1.5} + 1 = \frac{\ln\left(\frac{E}{2}\right)}{1.5 \ln 10} + 1. \tag{5.4}$$

We thus expect the energy to have a truncated Pareto-type distribution. Therefore, we apply the estimator of Beirlant et al. (2016a) to the energy and transform the endpoint back to the magnitudes using (5.4). Denote by Y_E the parent variable of E, which means that we observe $E =_d Y_E \,|\, Y_E < T_E$, with T_E the endpoint for E, see Table 5.1. The extreme value index of Y_E, ξ_{Y_E}, is estimated by solving (4.21) for ξ which gives the estimate $\hat{\xi}_k^{Y_E,+}$. Beirlant et al. (2016a) propose to do this using the Newton-Raphson algorithm.

The endpoint for the energy is estimated as

$$\hat{T}_k^{E,+} = 2 \times 10^{1.5(M_{n-k,n}-1)} \left(\frac{R_{k,n}^{1/\hat{\xi}_k^{Y_E,+}} - \frac{1}{k+1}}{1 - \frac{1}{k+1}} \right)^{-\hat{\xi}_k^{Y_E,+}} \tag{5.5}$$

with

$$R_{k,n} = \frac{2 \times 10^{1.5(M_{n-k,n}-1)}}{2 \times 10^{1.5(M_{n,n}-1)}} = 10^{1.5(M_{n-k,n}-M_{n,n})}.$$

Note that this estimator corresponds to (4.23) with $p = 0$. Transforming the estimated endpoint for the energy gives following endpoint estimate for the magnitudes:

$$\hat{T}_k^{M,+} = \frac{\log_{10}\left(\frac{\hat{T}_k^{E,+}}{2}\right)}{1.5} + 1. \tag{5.6}$$

We denote this estimator by *Truncated Pareto*.

Using the asymptotic results in Beirlant et al. (2016a), an approximate $100(1-\alpha)\%$ upper confidence bound for T_E can be constructed. Theorem 2 in Beirlant et al. (2016a) states that, after omitting second-order terms again,

$$\ln \hat{T}_k^{E,+} - \ln T_E = -\frac{\xi_{Y_E} \frac{k+1}{(n+1)D_T^E}}{k+1}(E_1 - 1)$$

where E_1 is a standard exponential random variable and D_T^E the truncation odds of E. Based on this result, we obtain following approximate $100(1-\alpha)\%$

upper confidence bound for T_E:

$$\exp\left(\ln \hat{T}_k^{E,+} - \frac{\frac{k+1}{(n+1)\hat{D}_{T,k}^{E,+}}\hat{\xi}_k^{Y_E,+}}{k+1}(\ln\alpha+1)\right),$$

where $\hat{D}_{T,k}^{E,+}$ is the truncated Pareto estimate for D_T^E. This upper bound can then be transformed back to the magnitude level as before to get an approximate $100(1-\alpha)\%$ upper confidence bound for T_M:

$$\frac{\ln\left(\frac{\hat{T}_k^{E,+}}{2}\right)}{1.5\ln 10}+1-\frac{\frac{\frac{k+1}{(n+1)\hat{D}_{T,k}^{E,+}}\hat{\xi}_k^{Y_E,+}}{k+1}(\ln\alpha+1)}{1.5\ln 10}=\hat{T}_k^{M,+}-\frac{\frac{\frac{k+1}{(n+1)\hat{D}_{T,k}^{E,+}}\hat{\xi}_k^{Y_E,+}}{k+1}(\ln\alpha+1)}{1.5\ln 10}. \tag{5.7}$$

Fraga Alves – Neves

Fraga Alves and Neves (2014) propose an endpoint estimator for distributions in the Gumbel MDA, i.e. $\xi_M = 0$. Fraga Alves et al. (2017) generalise this estimator to distributions in the Gumbel MDA and Weibull MDA with $\xi_M > -0.5$, i.e. $\xi_M \in (-0.5, 0]$. The generalised estimator is given by

$$\hat{T}_k^M = M_{n,n} + M_{n-k,n} - \frac{1}{\ln 2}\sum_{i=0}^{k-1}\ln\left(1+\frac{1}{k+i}\right)M_{n-k-i,n}, \tag{5.8}$$

for $k = 1, \ldots, \lfloor n/2 \rfloor$. We denote this estimator by *FAN*. Note that no estimator for ξ or ξ_M is used here.

Fraga Alves et al. (2017) also give an approximate $100(1-\alpha)\%$ upper confidence bound for T_M:

$$\hat{T}_k^M - N_{k,n}^{(1)}\left(1-\hat{\xi}_k^{M,-}\right)\left(h\left(\hat{\xi}_k^{M,-}\right)+k^{\hat{\xi}_k^{M,-}}\frac{(-\ln\alpha)^{-\hat{\xi}_k^{M,-}}}{\hat{\xi}_k^{M,-}}\right) \tag{5.9}$$

where $h(y) = \frac{1}{y}\left(\frac{2^{-y}-1}{y\ln 2}+1\right)$ and

$$\hat{\xi}_k^{M,-} = 1 - \frac{1}{2}\left(1-\frac{\left(N_k^{(1)}\right)^2}{N_k^{(2)}}\right)^{-1}$$

with $N_k^{(j)} = \frac{1}{k}\sum_{l=0}^{k-1}\left(X_{n-l,n} - X_{n-k,n}\right)^j$ for $j = 1, 2$ (Ferreira et al., 2003). This is a consistent estimator of ξ_M when $\xi_M < 0$.

As remarked in Section 4.2, an upper truncated distribution has EVI $\xi_M = -1$, but we still include the FAN estimator for comparison.

5.2.2 Non-parametric estimators

The next estimators are all based on the fact that

$$E(M_{n,n}) = \int_{t_M}^{T_M} m\, dF_M^n(m) = T_M - \int_{t_M}^{T_M} F_M^n(m)\, dm, \tag{5.10}$$

where F_M is the CDF of M, see Kijko and Singh (2011). Hence, T_M can be estimated by

$$\hat{T}^M = M_{n,n} + \Delta$$

with Δ an estimator for $\int_{t_M}^{T_M} F_M^n(m)\, dm$.

Non-parametric with Gaussian kernel

The CDF in (5.10) can be estimated using a Gaussian kernel. The estimator for the endpoint is then obtained as the iterative solution of the equation

$$T_M = M_{n,n} + \Delta \tag{5.11}$$

with

$$\Delta = \int_{t_M}^{T_M} \left(\frac{\sum_{i=1}^{n} \Phi\left(\frac{m-M_i}{h}\right) - \Phi\left(\frac{t_M-M_i}{h}\right)}{\sum_{i=1}^{n} \Phi\left(\frac{T_M-M_i}{h}\right) - \Phi\left(\frac{t_M-M_i}{h}\right)} \right)^n dm \tag{5.12}$$

and Φ the CDF of the standard normal distribution. The bandwidth h is chosen using unbiased cross-validation. We denote this estimator as *N-P-G*. For more details we refer to Kijko et al. (2001) and Equations 28 and 29 in Kijko and Singh (2011).

Non-parametric based on order statistics

Cooke (1979) proposes to approximate the CDF in (5.10) by the empirical CDF. The corresponding endpoint estimator, see Equation 33 in Kijko and Singh (2011), is given by

$$\hat{T}_n^M = M_{n,n} + \left[M_{n,n} - (1 - \exp(-1)) \sum_{i=0}^{n-1} \exp(-i) M_{n-i,n} \right]. \tag{5.13}$$

We denote this estimator as *N-P-OS*.

Cooke (1979) also constructed an approximate $100(1-\alpha)\%$ upper confidence bound for T_M:

$$M_{n,n} + \frac{M_{n,n} - M_{n-1,n}}{(1-\alpha)^{-\nu} - 1}, \tag{5.14}$$

where the parameter ν is determined by

$$\lim_{y\uparrow 0} \frac{1 - F_M(T_M + cy)}{1 - F_M(T_M + y)} = c^{1/\nu} \tag{5.15}$$

for every constant $c > 0$.

When a distribution is in the Weibull MDA, i.e. EVI $\xi_M < 0$, it has an upper bound. Notable examples are the uniform and beta distributions. As remarked in Section 4.2, an upper truncated distribution has EVI -1, and is hence in the Weibull MDA. However, the opposite is not true: not all distributions in the Weibull MDA are upper truncated as they do not always have an underlying parent distribution. For example, the beta distribution is bounded by 1 but is not truncated at 1. The GR distribution is bounded by T_M and has the (lower) truncated exponential distribution as parent distribution. Therefore it is upper truncated, and it thus is in the Weibull MDA with $\xi_M = -1$.

For a distribution with CDF F_M that is in the Weibull MDA one has

$$1 - F_M\left(T_M - \frac{1}{x}\right) = x^{\frac{1}{\xi_M}} \ell_{F_M}(x)$$

for $x \uparrow +\infty$, with ℓ_{F_M} a slowly varying function and T_M the endpoint of the distribution as before, see Section 2.4.2 in Beirlant et al. (2004). The left hand side of (5.15) then becomes

$$\lim_{y\uparrow 0} \frac{\left(-\frac{1}{cy}\right)^{\frac{1}{\xi_M}} \ell_{F_M}\left(-\frac{1}{cy}\right)}{\left(-\frac{1}{y}\right)^{\frac{1}{\xi_M}} \ell_{F_M}\left(-\frac{1}{y}\right)} = c^{-\frac{1}{\xi_M}} \lim_{y\uparrow 0} \frac{\ell_{F_M}\left(-\frac{1}{cy}\right)}{\ell_{F_M}\left(-\frac{1}{y}\right)} = c^{-\frac{1}{\xi_M}} \lim_{z\to+\infty} \frac{\ell_{F_M}\left(\frac{z}{c}\right)}{\ell_{F_M}(z)} = c^{-\frac{1}{\xi_M}}$$

where we used the definition of a slowly function (see page 49). This means that $\nu = -\frac{1}{\xi_M}$ in (5.15) for a distribution in the Weibull MDA. As upper truncated distributions are in the Weibull MDA with $\xi_M = -1$, they have $\nu = 1$. This is also remarked in Cooke (1979), but not proved there. Since it is often assumed that magnitude data come from an upper truncated distribution, e.g. the Gutenberg-Richter distribution, we use $\nu = 1$ in the remainder.

Few largest observations

Later, Cooke (1980) proposed a simple estimator that only uses the maximum and the $(k+1)$-th largest magnitude. This estimator, see Equation 38 in Kijko and Singh (2011), is equal to

$$\hat{T}_k^M = M_{n,n} + \left[\frac{1}{k}(M_{n,n} - M_{n-k+1,n})\right]. \tag{5.16}$$

We denote this estimator as *FL*.

Extended FL

The previous estimator only uses two observations. It can be extended as

$$\hat{T}_k^M = M_{n,n} + \left[\frac{1}{k}\left(M_{n,n} - \frac{1}{k-1}\sum_{i=2}^{k} M_{n-i+1,n}\right)\right], \tag{5.17}$$

see Equation 40 in Kijko and Singh (2011). We denote this estimator as *EFL*.

Robson–Whitlock

Robson and Whitlock (1964) propose the following simple estimator:

$$\hat{T}_2^M = M_{n,n} + [M_{n,n} - M_{n-1,n}], \tag{5.18}$$

see Equation 42 in Kijko and Singh (2011) We denote this estimator as *R-W*.

Another approximate $100(1-\alpha)\%$ upper confidence bound for T_M was derived by Robson and Whitlock (1964):

$$M_{n,n} + \frac{1-\alpha}{\alpha}(M_{n,n} - M_{n-1,n}). \tag{5.19}$$

Note that this corresponds to the upper confidence bound (5.14) of Cooke (1979) (with $\nu = 1$).

Robson–Whitlock–Cooke

The previous estimator can be improved, in terms of MSE, as shown in Cooke (1979). The improved estimator is obtained as

$$\hat{T}_2^M = M_{n,n} + \left[\frac{1}{2\nu}(M_{n,n} - M_{n-1,n})\right], \tag{5.20}$$

see Equation 46 in Kijko and Singh (2011). As before, we take ν equal to 1. We denote this estimator as *R-W-C*. Note that this estimator corresponds to the FL estimator for $k = 2$.

5.2.3 Parametric estimator: Kijko–Sellevol

Kijko and Sellevoll (1989) introduced the equation (see Equation 13 in Kijko and Singh (2011))

$$T_M = M_{n,n} + \left[\frac{E_1(n_2) - E_1(n_1)}{\beta \exp(-n_2)} + t_M \exp(-n) \right] \tag{5.21}$$

with

$$n_1 = \frac{n}{1 - \exp(-\beta(T_M - t_M))}, \quad n_2 = n_1 \exp(-\beta(T_M - t_M)),$$

and $E_1(z) = \int_z^\infty \exp(-s)/s \, ds$ the exponential integral function. Since these expressions depend on T_M, we obtain T_M using an iterative procedure. The parameter β is estimated using ML based on the Gutenberg-Richter law, see Page (1968) and Chapter 12 in Gibowicz and Kijko (1994). It is estimated iteratively using the equation

$$\frac{1}{\beta} = \overline{M}_n - t_M + \frac{(T_M - t_M)\exp(-\beta(T_M - t_M))}{1 - \exp(-\beta(T_M - t_M))},$$

where $\overline{M}_n = 1/n \sum_{i=1}^n M_i$ is the sample mean of $M_1, \ldots, M_n$. Using a Taylor expansion, this becomes

$$\hat{\beta} = \hat{\beta}_0 \left(1 - \hat{\beta}_0 \frac{(T_M - t_M)\exp(-\hat{\beta}_0(T_M - t_M))}{1 - \exp(-\hat{\beta}_0(T_M - t_M))} \right) \tag{5.22}$$

where $\hat{\beta}_0 = \frac{1}{\overline{M}_n - t_M}$ is the Aki-Utsu (Aki, 1965; Utsu, 1965) estimator for β. This approach does not use iterations and is thus preferred for computational reasons. In each iteration step (for T_M), we first update the estimate of β using (5.22), and then improve the estimate of T_M. We denote this estimator of the maximum magnitude as *K-S*. Note that this estimator is the only one that uses the Gutenberg-Richter law directly.

Based on the Gutenberg-Richter law, a parametric $100(1-\alpha)\%$ upper confidence bound for T_M can be constructed (Holschneider et al., 2011):

$$t_M - \frac{1}{\beta} \ln \left(\frac{\exp(-\beta(M_{n,n} - t_M)) - 1}{\alpha^{1/n}} + 1 \right), \tag{5.23}$$

where we estimate β using the K-S method. Holschneider et al. (2011) and Zöller and Holschneider (2016a) note that this upper bound is infinite if the maximum observed value is larger than $t_M - \frac{1}{\beta} \ln(1 - \alpha^{1/n})$. For the Groningen

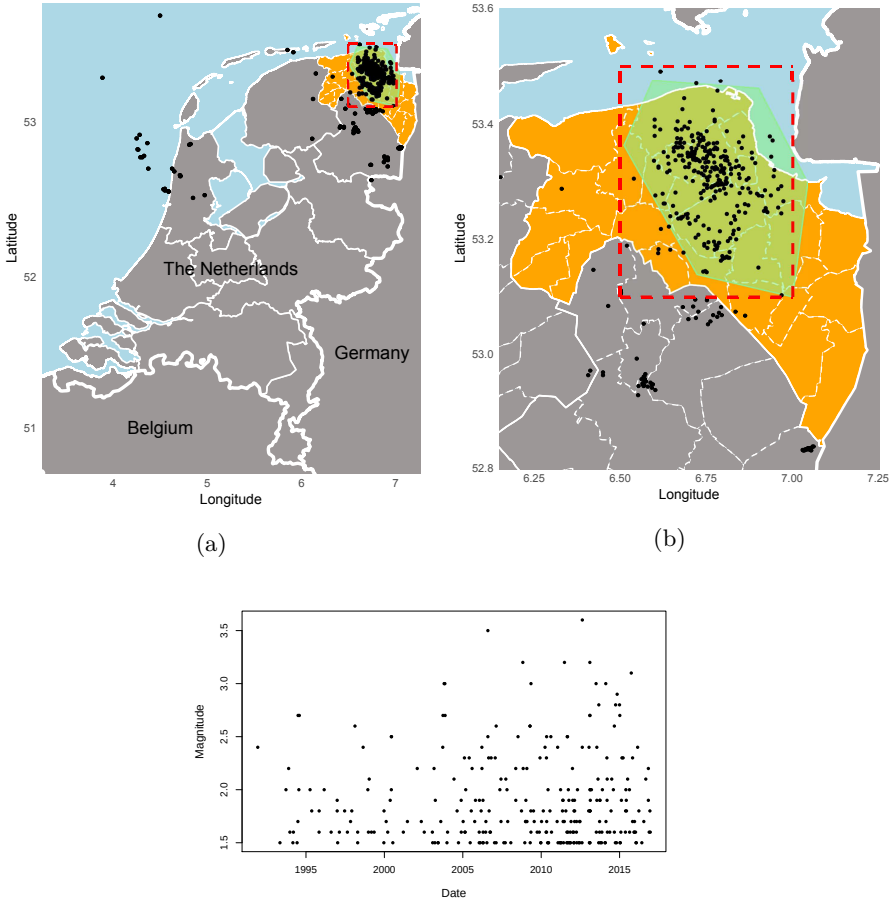
example, this happens when $\alpha \leq 0.061$. Therefore, we consider $\alpha = 0.1$ in the data example and the simulations.

Moreover, Holschneider et al. (2014) propose a statistical test for the maximum earthquake magnitude based on the GR distribution. To get a suitable testing power, however, an unrealistic large amount of large magnitudes is needed.

5.3 Estimation of the endpoint for Groningen

We now estimate the maximum possible earthquake magnitude in Groningen. We downloaded data on induced earthquakes in the Netherlands from the KNMI: `https://www.knmi.nl/kennis-en-datacentrum/dataset/aardbevingscatalogus`. The locations of the induced earthquakes with magnitudes larger than 1.5 are plotted in Figure 5.1a. As we are interested in the case of Groningen gas field, we consider the rectangle determined by (53.1°N, 6.5°E), (53.1°N, 7°E), (53.5°N, 7°E) and (53.5°N, 6.5°E). This is close to the area that was considered by Zöller and Holschneider (2016b). In this area, 286 earthquakes with magnitudes larger than 1.5 have been recorded between December 1986 and 31 December 2016. Their locations are plotted, together with the boundaries of the area, in Figure 5.1b. Moreover, the approximate location of the Groningen gas field is added in light green. The time plot of these earthquakes is given in Figure 5.1c. The dataset was tested for serial correlation and no significance could be detected.

Before applying the estimators, we smoothed the data by adding uniform noise $U[-0.05, 0.05]$ as the magnitudes are rounded up to one decimal digit. We then retain the 249 smoothed magnitudes larger than $t_M = 1.5$. The choice of 1.5 as threshold in the Groningen case is standard in the geological literature, see e.g. Dost et al. (2013). The exponential QQ-plot in Figure 5.2b indicates that an exponential distribution is indeed suitable for the magnitudes, but the bending off at the largest observations suggests an upper truncated tail. The same behaviour is seen on the mean excess plot (see e.g. Chapter 1 in Beirlant et al., 2004) in Figure 5.2c: the first horizontal part suggests that the data come from an exponential-like distribution, whereas the downward trend at the end indicates an upper truncation point. Note that the Pareto QQ-plot of the energy in Figure 5.2a suggests that the energy follows a truncated Pareto distribution as discussed in Section 5.2.1. When applying the truncated GPD estimator to the magnitudes, a value of ξ around 0 is found suggesting again an exponential-like distribution, see Figure 5.3a. The parameter ξ_{Y_E} is estimated by the truncated Pareto estimator to be around 1.8. The estimators for D_T based on the previous estimators for ξ suggest a truncation odds around 1%,



(a) (b)

(c)

Figure 5.1: Locations of induced earthquakes in (a) the Netherlands and (b) Groningen between December 1986 and 31 December 2016 with magnitudes larger than 1.5, and (c) time plot of induced earthquakes in Groningen with magnitudes larger than 1.5 in the considered area.

see Figure 5.3b. Moreover, the test for truncation based on the truncated GPD in Figure 5.3c, indicates, for larger values of k, that the data come indeed from an upper truncated distribution. Finally, the fit provided by the truncated GPD with $k = 150$, and hence $\hat{\xi}_{150} \approx 0$, models the data well, see Figure 5.2d. All these observations suggest that the magnitude data come indeed from the Gutenberg-Richter distribution, i.e. a doubly truncated exponential distribution.

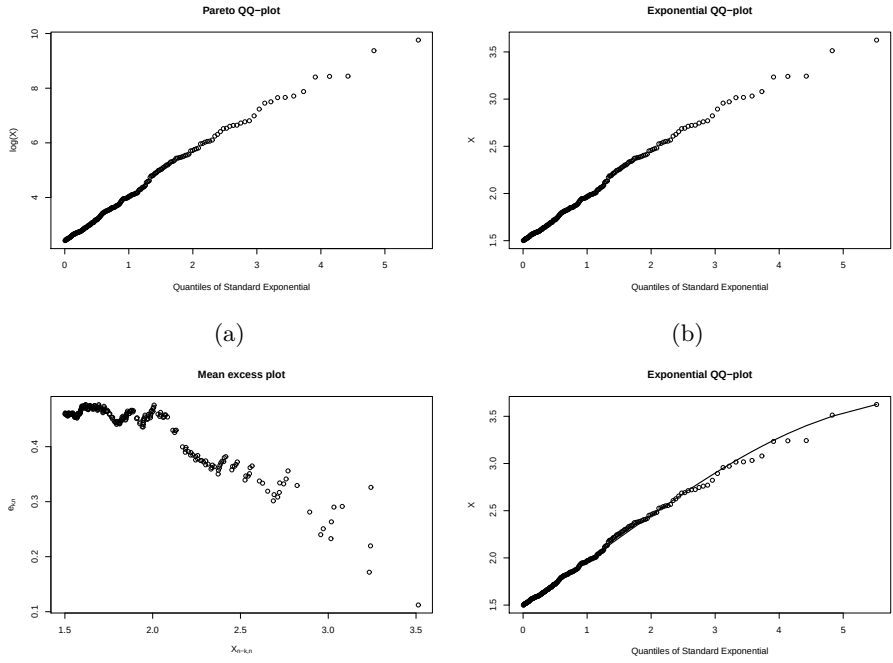
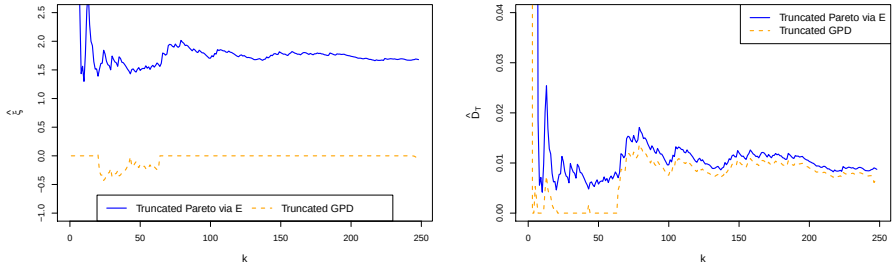


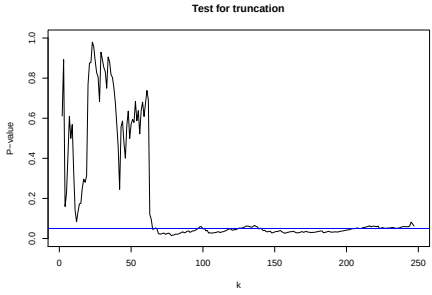
Figure 5.2: Groningen earthquakes: (a) Pareto QQ-plot of energy data, (b) exponential QQ-plot of magnitude data (c) mean excess plot of magnitude data and (d) exponential QQ-plot of magnitude data with fit based on $k = 150$ largest magnitudes.

Next, we compute all discussed estimates for the maximum possible earthquake magnitude (Figure 5.4a). For estimators that do not depend on k, the dot indicates how many observations are used: 2 or n. All estimators suggest that the endpoint lies between 3.6 and 3.9 on the Richter scale. Note however, that for the estimators of the endpoint based on EVT, we need to look at larger values of k where a more stable pattern emerges as the test for truncation was only significant for $k \geq 70$. For k around 150, the EVT methods based on truncation estimate the endpoint around 3.8. Note that the EVT estimates for $k = n$ are close to the estimates of the N-P-G and K-S methods which use all n observations. All other methods lead to lower estimates for the endpoint than the EVT methods.



(a)

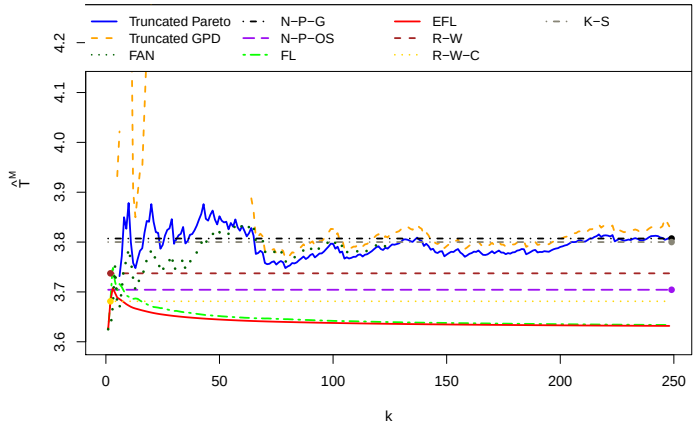
(b)



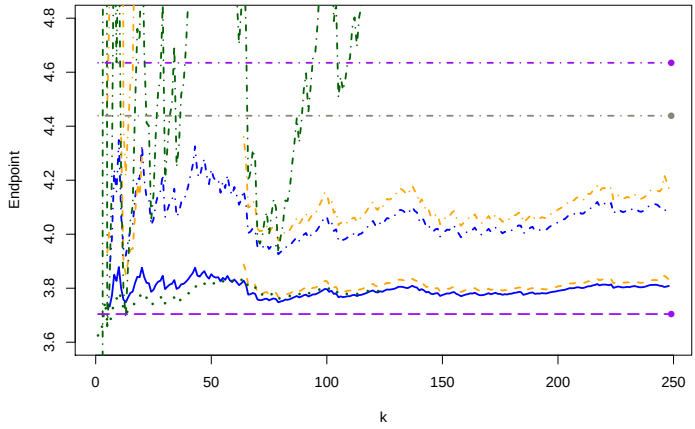
(c)

Figure 5.3: Groningen earthquakes: (a) estimates of ξ, (b) estimates of the truncation odds D_T and (c) test for truncation.

Additionally, we look at 90% upper confidence bounds for the endpoint as discussed above. The endpoint estimators are given by the full blue (truncated Pareto), dashed orange (truncated GPD), dotted green (FAN) and purple long dashed (N-P-OS) lines in Figure 5.4b. The corresponding 90% upper bounds are added as dash-dotted lines in the same colour. The upper bounds using the truncated Pareto (5.7) and truncated GPD (5.2) take values of 4 and 4.05, respectively, for $k = 150$. The 90% upper bound (5.14) takes a value of 4.63, and the parametric 90% upper bound (5.23) is equal to 4.44 (grey point). Note that the latter two confidence bounds are based on n magnitudes and should hence be compared with the upper bounds using truncated EVT for $k = n$ (4.10 and 4.17, respectively). The upper bound based on the FAN estimator (5.9) is rather volatile and takes values between 3.65 and 5.84. For example for $k = 80$, the endpoint is estimated as 3.76 and the upper bound is 3.99 which is in line with the results from the other two EVT-based estimators.



(a)



(b)

Figure 5.4: Groningen earthquakes: (a) estimates of the maximum possible magnitude T_M and (b) 90% upper confidence bounds for T_M.

5.4 Simulations

We now perform a simulation study based on the data example. We generate 1000 samples of size 250 from the Gutenberg-Richter distribution with $t_M = 1.5$ and $\beta = 2.1151$. The parameter β was estimated by the K-S estimator on the Groningen data. We consider different simulations from a Gutenberg-Richter distribution with these parameters where we let the endpoint vary: 3.75, 4 and 4.5. Note that these endpoints correspond to the 99.1%, 99.5% and 99.8% quantiles of the exponential distribution with rate 2.1151 and lower truncation point $t_M = 1.5$. For each of these simulations, we plot the relative mean, the relative MSE and the coverage percentage of the upper confidence bounds over the 1000 simulations. These plots can be found in Appendix C.

We see that the truncated Pareto and truncated GPD estimators have the lowest bias, over all truncation points. However, their MSE is among the highest which indicates that these estimators have a larger variance than the traditional endpoint estimators. As expected, the bias and MSE of the estimators increases when the endpoint gets larger. When simulating from the Gutenberg-Richter distribution with an endpoint of 3.75 or 4, which seems to be realistic based on the Groningen data, the truncated EVT estimators overestimate the true endpoint, on average. When $T_M = 4.5$, the estimates are, on average, too low. All other estimates, except the parametric K-S estimator, are on average always too low. The FAN estimator has very low bias when $T_M = 3.75$, but this becomes much worse than the two other EVT-based methods when the endpoint is larger. However, it has the lowest MSE across all methods and endpoints.

The coverage percentages of the upper confidence bounds are defined as the percentage of times that the obtained upper bounds are larger than the true endpoint. In theory these percentages should be equal to 90%. When the endpoint gets larger, the observed coverage percentages decrease. The coverage percentage for the upper bound of Cooke (1979) is closer to 90% than the ones for the upper bounds of the EVT-based estimators. The performance of the two first EVT-based upper bounds is rather similar with a slight advantage for the truncated Pareto. Since second-order bias terms were not taken into account for the upper bounds (5.2) and (5.7), developing bias reduced methods can improve these upper bounds. The upper confidence bound based on the FAN estimator (5.9) performs well, even for higher endpoints, although the upper bound is on average too high when $T_M = 3.75$. Note that this upper bound takes second order terms into account which leads to better coverage percentages than for the other two EVT-based upper confidence bounds. Although the upper bound performs well on average, for a single dataset the same volatile behaviour as in the Groningen example can be seen making it difficult to select a suitable value

for k. The parametric upper confidence bound (5.23), which uses n observations, performs similar to the one using the truncated Pareto for k large when the endpoint is 3.75. For higher endpoints, this upper confidence bound performs much worse than the other ones.

5.5 Conclusions

As an application of the methods from the previous chapter, we looked at the estimation of the maximum possible earthquake magnitude. We also made the comparison with other EVT-based estimators for the endpoint, with non-parametric methods from the geophysical literature and with a parametric method based on the Gutenberg-Richter distribution. Since there are only a few large earthquakes, there is a lot of uncertainty when estimating the maximum possible earthquake, even for the methods from EVT. Therefore, it is important to quantify the uncertainty using confidence bounds. Zöller and Holschneider (2016a) note that using additional information, apart from the magnitude data, can make matters worse since there is also uncertainty on this information.

Using the considered methods, the maximum possible magnitude in Groningen is estimated to be in the range 3.65 to 3.9. 90% upper confidence bounds based on these methods vary from 4 to 4.65. Moreover, our extreme value analysis also indicates that the widely used Gutenberg-Richter distribution is indeed appropriate to model the earthquake magnitudes in Groningen. However, the EVT-based and non-parametric estimators do not use this distribution which gives them more flexibility compared to parametric estimators.

Based on simulations from the GR distribution, it is clear that the EVT-based methods of the previous chapter and Beirlant et al. (2016a) perform well when estimating the endpoint. It is important to note that these methods usually have positive bias which means that they, on average, overestimate the true endpoint, whereas the other estimators are too low, on average. The upper confidence bounds based on these two estimators are sharper than the other ones, however, the simulations point out that they are too sharp indicating the need for bias reduction.

Overall, we can conclude that the EVT-based estimators of the previous chapter and Beirlant et al. (2016a) are a valuable addition to the existing methods for estimating the maximum possible earthquake magnitude.

Chapter 6

Modelling censored losses using splicing: a global fit strategy with mixed Erlang and extreme value distributions

This chapter is based on

Reynkens, T., Verbelen, R., Beirlant, J. and Antonio, K. (2017). Modelling Censored Losses Using Splicing: a Global Fit Strategy With Mixed Erlang and Extreme Value Distributions, available on arXiv:1608.01566.

6.1 Introduction

In several domains such as insurance, finance and operational risk, modelling financial losses is essential. For example, actuaries use models for claim sizes to set premiums, calculate risk measures and determine capital requirements for solvency regulations. This type of data is typically heavy-tailed and high losses can occur. A standard parametric distribution for the tail is a Pareto-type distribution, which is of key importance in extreme value theory (see the

introduction of Chapter 3 and e.g. McNeil, 1997). The Pareto distribution or the GPD are used to model exceedances over intermediate thresholds. However, they are not able to capture the characteristics over the whole range of the loss distribution which makes them not suitable as a global fit distribution. It is often imperative to obtain a global fit for the distribution of losses, for example in a risk analysis where focus is not only on extreme events, or when setting up a reinsurance program. Instead of trying many different standard distributions, splicing two distributions (Klugman et al., 2012) is more suitable to model the complete loss distribution. In literature, a splicing model is also called a composite model. We hereby combine a light-tailed distribution for the body which covers light and moderate losses (the so-called attritional losses), and a heavy-tailed distribution for the tail to capture large losses. In the actuarial literature simple splicing models have been proposed. Beirlant et al. (2004) and Klugman et al. (2012) consider the splicing of the exponential distribution with the Pareto distribution. Other distributions for the body such as the Weibull distribution (Ciumara, 2006; Scollnik and Sun, 2012) or the lognormal distribution (Cooray and Ananda, 2005; Scollnik, 2007; Pigeon and Denuit, 2011) have also been used. Nadarajah and Bakar (2014), Bakar et al. (2015) and Calderín–Ojeda and Kwok (2016) investigate the splicing of the lognormal or Weibull distribution with various tail distributions. Lee et al. (2012) consider the splicing of a mixture of two exponentials and the GPD. The use of a mixture model in the first splicing component gives more flexibility in modelling the light and moderate losses. Fackler (2013) provides an overview of spliced distributions for loss modelling. Note that splicing has not only been considered in an actuarial context. Panjer (2006), Peters and Shevchenko (2015) and Aue and Kalkbrener (2006) use this technique to model operational risk data.

The mixed Erlang (ME) distribution became popular in loss modelling because of several reasons (see e.g. Willmot and Woo, 2007; Lee and Lin, 2010; Willmot and Lin, 2011; Klugman et al., 2013). The class of ME distributions with common scale parameter is dense in the space of positive continuous distributions (Tijms, 1994). Any positive continuous distribution can thus be approximated up to any given accuracy by a ME distribution. This class is also closed under mixture, convolution and compounding. Therefore, we can readily obtain aggregate loss distributions removing the need for simulations. Moreover, we can easily compute risk measures such as the VaR, the Tail Value-at-Risk (TVaR) and premiums of excess-loss insurances.

Fitting the ME distribution using direct likelihood maximisation is difficult. The ME parameters can also be estimated based on the denseness proof of Tijms (1994) but this method converges slowly and leads to overfitting (Lee and Lin, 2010). The preferred strategy is to use the expectation-maximisation

(EM) algorithm (Dempster et al., 1977) to fit the ME distribution as proposed by Lee and Lin (2010). An advantage is that the E- and M-steps can be solved analytically. Lee and Lin (2010) use information criteria (IC) like the Akaike information criterion (AIC, Akaike, 1974) or the Bayesian information criterion (BIC, Schwarz, 1978) to select the number of components in the mixture and as such avoid overfitting.

Our work is further inspired by the omnipresence of censoring and truncation in risk analysis and risk modelling, see e.g. Cao et al. (2009), Klugman et al. (2012), Antonio and Plat (2014) and Verbelen et al. (2015).

Lower truncation occurs when payments that are below certain thresholds are not observed. In insurance, lower truncation occurs, for example, due to the presence of a deductible in the insurance contract. In some practical applications, there might be a natural bound that upper truncates the tail distribution. For example, earthquake magnitudes (as seen in the previous chapter) and forest fire areas have distributions that are naturally *upper truncated* (Beirlant et al., 2016a). In an insurance context, where premiums have to be set using the fitted model, introducing an upper truncation point can prevent probability mass being assigned to unreasonably large claim amounts.

Right censoring is highly relevant in the context of loss models and risk measurement for unsettled claims in non-life insurance and reinsurance. The (re)insurer only knows the true cost of a policy when all claims on this policy are settled or closed. However, in the development or lifetime of a non-life insurance claim, a significant time may elapse between the claim occurrence and its final settlement or closure. For such unsettled claims only the payment to date is known and the quantity of interest, i.e. the final cumulative payment on a claim, is right censored. This complicates the calculation of reinsurance premiums for large claims and forces the insurer to predict, with maximum accuracy, the capital buffer that is required to indemnify the insured in the future regarding claims that happened in the past. To support this complex task, actuaries will use additional, expert information called *incurred data*. This is the sum of the actual payment (so far) on a claim and its case estimate. These case estimates are set by an experienced case handler and express the expert's estimate of the outstanding loss on a claim. For large claims, facing very long settlement (e.g. due to legal procedures or severe bodily injury), actuaries consider incurred data as a highly important source of information. We propose to construct upper bounds for the final cumulative payment on a claim using incurreds. When the true final claim amount lies between the cumulative payment up to date and the incurred value, *interval censoring* techniques can be applied as we will demonstrate later in this chapter.

The previous example covers random censoring. Policy limits introduce another type of censoring, namely type I (right) censoring (see e.g. Klein and Moeschberger, 2003). When the loss corresponding to a claim exceeds the policy limit, no further payments need to be made by the insurer. The loss is thus censored and only known to be larger than the amount that needs to be paid by the insurer, i.e. the policy limit. In the remainder, we only consider random censoring and not type I censoring.

In the splicing context, some work has already been done for censored and/or truncated data. Teodorescu and Panaitescu (2009) take lower truncation into account for Weibull-Pareto splicing, and Cooray and Ananda (2005) extend their approach to type I right censored data. Verbelen et al. (2015) extend the mixed Erlang approach of Lee and Lin (2010) to censored and/or truncated data. Beirlant et al. (2007) and Einmahl et al. (2008) discuss extensions of classical extreme value estimators to right censored data. Extensions to upper truncated data have been investigated by Aban et al. (2006), Beirlant et al. (2016a) and Beirlant et al. (2017), see Chapter 4.

Although the ME distribution has several advantages, as discussed above, one major disadvantage is that it has an asymptotically exponential, and hence: light, tail (Neuts, 1981). Therefore, overfitting can still occur on heavy-tailed data as one needs many components to model the heavy-tailedness appropriately. The simulated sample of the GPD in Verbelen et al. (2015) illustrates this behaviour. As a first contribution, we overcome this drawback by proposing a splicing model with the ME distribution for the body and the Pareto distribution for the tail (Section 6.2). A global fit for financial loss data then results, which combines the flexibility of the ME distribution to model light and moderate losses with the ability of the Pareto distribution to model heavy-tailed data. Fire and motor third party liability (MTPL) insurance losses, and financial returns are examples of heavy-tailed data which are of Pareto type. This strategy avoids ad hoc combinations of a standard light-tailed distribution, such as the lognormal or the Weibull distribution, for the body with a heavy-tailed distribution for the tail, as explored in many papers on loss modelling. Moreover, a mixture of Erlangs yields more flexibility than a mixture of two exponentials as in Lee et al. (2012) while keeping analytic tractability.

As a second contribution, we extend the global fit strategy based on splicing to take both (random) censoring and truncation into account. Up to our knowledge, this full framework has not yet been considered in the literature. We provide a general fitting procedure for the model using the EM algorithm where the incompleteness is caused by censoring, see Section 6.3. Instead of using a splicing model, a common technique in extreme value analysis is to combine a non-parametric fit for the body and a parametric model (e.g. Pareto

distribution) for the tail. However, when censoring is present, this approach can no longer be applied as we might have interval censored data points where the lower bound of the interval is in the body of the distribution, whereas the upper bound is in the tail. Our general splicing framework can handle observations of this type and can hence be used to provide a global fit. As we provide a general procedure to fit a splicing model to censored and/or truncated data, we could possibly use another extreme value distribution, such as the GPD, instead of the Pareto distribution. For the GPD, however, in case there is censoring, the expectations in the E-step can no longer be computed analytically, in contrast to the Pareto distribution.

In Section 6.4, we apply the general fitting procedure for censored and/or truncated data to the specific case of our ME-Pareto splicing model. The incompleteness now stems on the one hand from censoring and on the other hand from the mixing of Erlang components. The general fitting procedure is therefore extended using ideas from the procedure of Verbelen et al. (2015) for fitting the ME distribution to censored and/or truncated data.

Finally, we discuss the computation of risk measures using our splicing model in Section 6.5 and we apply the method to two real life data examples in Section 6.6.

6.2 Splicing of ME and Pareto distributions

6.2.1 General splicing model

Consider two densities f_1^* and f_2^*, and denote the corresponding CDFs by F_1^* and F_2^*. Their parameters are contained in the vectors $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$, respectively. We assume that there are no shared parameters in $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$. Define now

$$f_1(x; t^l, t, \boldsymbol{\Theta}_1) = \begin{cases} \frac{f_1^*(x;\boldsymbol{\Theta}_1)}{F_1^*(t;\boldsymbol{\Theta}_1) - F_1^*(t^l;\boldsymbol{\Theta}_1)} & \text{if } t^l \leq x \leq t \\ 0 & \text{otherwise,} \end{cases}$$

$$f_2(x; t, T, \boldsymbol{\Theta}_2) = \begin{cases} \frac{f_2^*(x;\boldsymbol{\Theta}_2)}{F_2^*(T;\boldsymbol{\Theta}_2) - F_2^*(t;\boldsymbol{\Theta}_2)} & \text{if } t \leq x \leq T \\ 0 & \text{otherwise,} \end{cases}$$

where $0 \leq t^l < t < T$ are fixed points. The first density is lower truncated at t^l and upper truncated at t, and the second density is lower truncated at t and upper truncated at T. The density for the body, f_1, and density for the tail, f_2, are then valid densities on the intervals $[t^l, t]$ and $[t, T]$, respectively.

In case of no upper truncation for the tail distribution, we set $T = +\infty$. The corresponding CDFs are

$$F_1(x; t^l, t, \boldsymbol{\Theta}_1) = \begin{cases} 0 & \text{if } x \leq t^l \\ \frac{F_1^*(x;\boldsymbol{\Theta}_1) - F_1^*(t^l;\boldsymbol{\Theta}_1)}{F_1^*(t;\boldsymbol{\Theta}_1) - F_1^*(t^l;\boldsymbol{\Theta}_1)} & \text{if } t^l < x < t \\ 1 & \text{if } x \geq t, \end{cases}$$

$$F_2(x; t, T, \boldsymbol{\Theta}_2) = \begin{cases} 0 & \text{if } x \leq t \\ \frac{F_2^*(x;\boldsymbol{\Theta}_2) - F_2^*(t;\boldsymbol{\Theta}_2)}{F_2^*(T;\boldsymbol{\Theta}_2) - F_2^*(t;\boldsymbol{\Theta}_2)} & \text{if } t < x < T \\ 1 & \text{if } x \geq T. \end{cases}$$

Consider the splicing weight $\pi \in (0, 1)$. The spliced density is then defined as

$$f(x; t^l, t, T, \boldsymbol{\Theta}) = \begin{cases} 0 & \text{if } x \leq t^l \\ \pi f_1(x; t^l, t, \boldsymbol{\Theta}_1) & \text{if } t^l < x \leq t \\ (1-\pi) f_2(x; t, T, \boldsymbol{\Theta}_2) & \text{if } t < x < T \\ 0 & \text{if } x \geq T, \end{cases}$$

where $\boldsymbol{\Theta} = (\pi, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ is the parameter vector. We call the point t the splicing point, and the points t^l and T the lower, respectively, upper truncation points. The corresponding, continuous, CDF is given by

$$F(x; t^l, t, T, \boldsymbol{\Theta}) = \begin{cases} 0 & \text{if } x \leq t^l \\ \pi F_1(x; t^l, t, \boldsymbol{\Theta}_1) & \text{if } t^l < x \leq t \\ \pi + (1-\pi) F_2(x; t, T, \boldsymbol{\Theta}_2) & \text{if } t < x < T \\ 1 & \text{if } x \geq T. \end{cases} \tag{6.1}$$

Most authors impose differentiability of the probability density function (PDF) at the splicing point to get a smooth density function and to reduce the number of parameters. The splicing point is then estimated together with the other model parameters using maximum likelihood estimation (MLE). This restriction results in less flexibility. Therefore, we choose to not follow this approach, but determine the splicing point directly using an extreme value analysis, see Section 6.4.3.

6.2.2 Mixed Erlang distribution

In our specific case, f_1 is the density of a mixed Erlang (ME) distribution which is lower truncated at $t^l \geq 0$ and upper truncated at $t > t^l$. More specifically,

we consider a mixture of M Erlang distributions with common scale parameter $\theta > 0$.

The Erlang distribution is a Gamma distribution with an integer shape parameter. It has density function

$$f_E(x; r, \theta) = \frac{x^{r-1} \exp(-x/\theta)}{\theta^r (r-1)!} \qquad \text{for } x > 0\,, \tag{6.2}$$

where r, a positive integer, is the shape parameter, and $\theta > 0$ is the scale parameter. Its inverse $\lambda = 1/\theta$ is called the rate parameter. Integrating (6.2) r times by parts gives the cumulative distribution function

$$F_E(x; r, \theta) = \int_0^x \frac{z^{r-1} \exp(-z/\theta)}{\theta^r (r-1)!} dz = 1 - \sum_{z=0}^{r-1} \exp(-x/\theta) \frac{(x/\theta)^z}{z!}\,.$$

The density of the ME distribution is then given by

$$f_1^*(x; \boldsymbol{\alpha}, \boldsymbol{r}, \theta) = \sum_{j=1}^{M} \alpha_j \frac{x^{r_j-1} \exp(-x/\theta)}{\theta^{r_j} (r_j-1)!} = \sum_{j=1}^{M} \alpha_j f_E(x; r_j, \theta) \qquad \text{for } x > 0\,,$$

where the positive integers $\boldsymbol{r} = (r_1, \ldots, r_M)$ with $r_1 < \ldots < r_M$ are the shape parameters of the Erlang distributions, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$, with $\alpha_j > 0$ and $\sum_{j=1}^{M} \alpha_j = 1$, are the mixing weights. Similarly, the cumulative distribution function can be written, for $x > 0$, as

$$F_1^*(x; \boldsymbol{\alpha}, \boldsymbol{r}, \theta) = \sum_{j=1}^{M} \alpha_j \left(1 - \sum_{z=0}^{r_j-1} \exp(-x/\theta) \frac{(x/\theta)^z}{z!} \right) = \sum_{j=1}^{M} \alpha_j F_E(x; r_j, \theta).$$

After truncation, with limits t^l and t, the probability density function becomes

$$f_1(x; t^l, t, \boldsymbol{r}, \boldsymbol{\Theta}_1) = \begin{cases} \dfrac{f_1^*(x; \boldsymbol{r}, \boldsymbol{\Theta}_1^*)}{F_1^*(t; \boldsymbol{r}, \boldsymbol{\Theta}_1^*) - F_1^*(t^l; \boldsymbol{r}, \boldsymbol{\Theta}_1^*)} \\ \qquad = \sum_{j=1}^{M} \beta_j f_E^t(x; t^l, t, r_j, \theta) & \text{for } t^l \le x \le t \\ 0 & \text{otherwise,} \end{cases}$$

with $\boldsymbol{\Theta}_1 = (\boldsymbol{\beta}, \theta)$, which is again a mixture with mixing weights

$$\beta_j = \alpha_j \frac{F_E(t; r_j, \theta) - F_E(t^l; r_j, \theta)}{F_1^*(t; \boldsymbol{r}, \boldsymbol{\Theta}_1^*) - F_1^*(t^l; \boldsymbol{r}, \boldsymbol{\Theta}_1^*)} \tag{6.3}$$

and component density functions

$$f_E^t(x; t^l, t, r_j, \theta) = \frac{f_E(x; r_j, \theta)}{F_E(t; r_j, \theta) - F_E(t^l; r_j, \theta)}\,.$$

The component density functions $f_E^t(x; t^l, t, r_j, \theta)$ are truncated versions of the original component density functions $f_E(x; r_j, \theta)$. We obtain the weights β_j by reweighting the original weights α_j using the probability of the corresponding mixing component to lie in the truncation interval. Denote by F_E^t the CDF corresponding to f_E^t. The CDF corresponding to f_1 is then given by

$$F_1(x; t^l, t, \boldsymbol{r}, \boldsymbol{\Theta}_1) = \begin{cases} 0 & \text{if } x \leq t^l \\ \sum_{j=1}^M \beta_j F_E^t(x; t^l, t, r_j, \theta) & \\ \quad = \sum_{j=1}^M \beta_j \frac{F_E(x;r_j,\theta) - F_E(t^l;r_j,\theta)}{F_E(t;r_j,\theta) - F_E(t^l;r_j,\theta)} & \text{if } t^l < x < t \\ 1 & \text{if } x \geq t. \end{cases} \tag{6.4}$$

The number of Erlang mixtures M and the positive integer shapes $\boldsymbol{r}$ are fixed when estimating $\boldsymbol{\Theta}_1 = (\boldsymbol{\beta}, \theta)$. They are chosen using the approach described in Section 4 of Verbelen et al. (2016). A short overview of this approach is included in Appendix D.1.4.

6.2.3 Pareto distribution

The second density f_2 is the density of the truncated Pareto distribution with scale parameter $t > 0$, shape parameter $\xi > 0$ and upper truncation point T that can be $+\infty$. Note that the scale parameter t coincides with the fixed lower truncation point of the tail distribution. As mentioned before, we determine it in advance using an extreme value analysis, see Section 6.4.3. Hence, $\boldsymbol{\Theta}_2 = \xi$. More precisely, we have

$$f_2(x; t, T, \xi) = \frac{f_2^*(x; t, \xi)}{F_2^*(T; t, \xi)} = \begin{cases} \frac{\frac{1}{\xi t}\left(\frac{x}{t}\right)^{-\frac{1}{\xi}-1}}{1-\left(\frac{T}{t}\right)^{-\frac{1}{\xi}}} & \text{if } t < x < T \\ 0 & \text{otherwise,} \end{cases}$$

and

$$F_2(x; t, T, \xi) = \begin{cases} 0 & \text{if } x \leq t \\ \frac{1-\left(\frac{x}{t}\right)^{-\frac{1}{\xi}}}{1-\left(\frac{T}{t}\right)^{-\frac{1}{\xi}}} & \text{if } t < x < T \\ 1 & \text{if } x \geq T. \end{cases} \tag{6.5}$$

6.3 Fitting a general splicing model to censored data using the EM algorithm

In this section, we discuss maximum likelihood estimation for fitting a general splicing model, as proposed in Section 6.2.1, to censored data. The special case of a splicing model that combines a mixed Erlang distribution (as introduced in Section 6.2.2) and a Pareto distribution (Section 6.2.3) is treated in the subsequent section. The parameters to be estimated are contained in the vector $\boldsymbol{\Theta} = (\pi, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$.

6.3.1 Randomly censored data

We represent the censored sample by $\mathcal{X} = \{(l_i, u_i) \,|\, i = 1, \ldots, n\}$, where l_i and u_i denote the lower and upper censoring points of each data point from the sample of size n. These censoring points must be interpreted as the lower and upper endpoints of the interval that contains the data point x_i, which is not always observed. The censoring status of each data point is determined as follows:

Uncensored:	$t^l \leq l_i = x_i = u_i \leq T$
Left censored:	$t^l = l_i < u_i \leq T$
Right censored:	$t^l \leq l_i < u_i = T$
Interval censored:	$t^l \leq l_i < u_i \leq T$.

The left censored and right censored data points can be treated as a special case of interval censored data points with $l_i = t^l$ and $u_i = T$, respectively. In the splicing context, we make a distinction between five cases of data points:

i. Uncensored with $t^l \leq l_i = x_i = u_i \leq t < T$

ii. Uncensored with $t^l < t < l_i = x_i = u_i \leq T$

iii. Interval censored with $t^l \leq l_i < u_i \leq t < T$

iv. Interval censored with $t^l < t \leq l_i < u_i \leq T$

v. Interval censored with $t^l \leq l_i < t < u_i \leq T$.

These cases are visualised in Figure 6.1. In case v, we make a further subdivision based on whether the unobserved data point lies below or above the splicing point t.

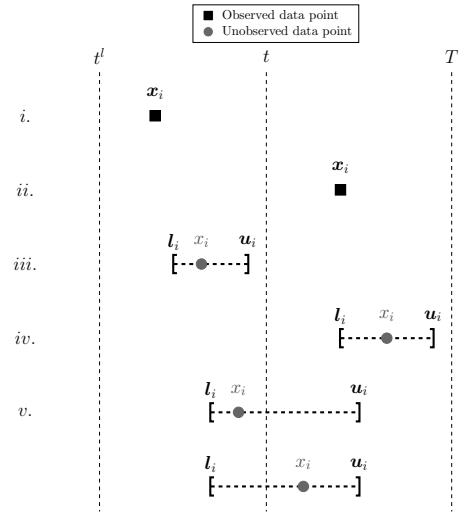


Figure 6.1: The different cases of data points.

6.3.2 Maximum likelihood estimation using the EM algorithm

We use maximum likelihood to fit the parameters of the spliced distribution. The likelihood function of the parameter vector $\boldsymbol{\Theta}$ is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Theta};\mathcal{X}) = &\prod_{i\in S_{i.}} \pi f_1(x_i;t^l,t,\boldsymbol{\Theta}_1) \prod_{i\in S_{ii.}} (1-\pi) f_2(x_i;t,T,\boldsymbol{\Theta}_2) \\ &\prod_{i\in S_{iii.}} \pi \left(F_1(u_i;t^l,t,\boldsymbol{\Theta}_1) - F_1(l_i;t^l,t,\boldsymbol{\Theta}_1)\right) \\ &\prod_{i\in S_{iv.}} (1-\pi) \left(F_2(u_i;t,T,\boldsymbol{\Theta}_2) - F_2(l_i;t,T,\boldsymbol{\Theta}_2)\right) \\ &\prod_{i\in S_{v.}} \left(\pi + (1-\pi)F_2(u_i;t,T,\boldsymbol{\Theta}_2) - \pi F_1(l_i;t^l,t,\boldsymbol{\Theta}_1)\right),\end{aligned}$$

where $S_{i.}$ is the subset of $\{1,\ldots,n\}$ corresponding to data points of case i, and similarly for the other cases. The corresponding log-likelihood is

$$\begin{aligned}\ell(\boldsymbol{\Theta};\mathcal{X}) = &\sum_{i\in S_{i.}} \left(\ln\pi + \ln f_1(x_i;t^l,t,\boldsymbol{\Theta}_1)\right) + \sum_{i\in S_{ii.}} \Big(\ln(1-\pi) + \ln f_2(x_i;t,T,\boldsymbol{\Theta}_2)\Big) \\ &+ \sum_{i\in S_{iii.}} \Big(\ln\pi + \ln\left(F_1(u_i;t^l,t,\boldsymbol{\Theta}_1) - F_1(l_i;t^l,t,\boldsymbol{\Theta}_1)\right)\Big)\end{aligned}$$

$$+ \sum_{i \in S_{iv.}} \Big(\ln(1-\pi) + \ln\Big(F_2(u_i; t, T, \boldsymbol{\Theta}_2) - F_2(l_i; t, T, \boldsymbol{\Theta}_2) \Big) \Big)$$

$$+ \sum_{i \in S_{v.}} \ln\Big(\pi + (1-\pi) F_2(u_i; t, T, \boldsymbol{\Theta}_2) - \pi F_1(l_i; t^l, t, \boldsymbol{\Theta}_1) \Big). \tag{6.6}$$

Direct numerical optimisation of the log-likelihood expression (6.6) is not straightforward due to the censoring. Data points corresponding to case v, where the censoring interval contains the splicing point t, lead to logarithmic terms of a sum involving the splicing weight π, the parameters $\boldsymbol{\Theta}_1$ of the body distribution as well as the parameters $\boldsymbol{\Theta}_2$ of the tail distribution of the splicing model. This prevents separate optimisation with respect to each of these parameter blocks.

We use the EM algorithm to overcome this hurdle in fitting a splicing model to censored data. This iterative method, first introduced by Dempster et al. (1977), finds the maximum likelihood estimates when the data are incomplete and direct likelihood maximisation is not easy to perform numerically. Consider the complete data $\mathcal{Y}$ containing the uncensored sample $\boldsymbol{x} = (x_1, \ldots, x_n)$. Given the complete version of the data, we can construct a complete likelihood function as

$$\mathcal{L}_{\text{complete}}(\boldsymbol{\Theta}; \mathcal{Y}) = \prod_{i=1}^{n} \big(\pi f_1(x_i; t^l, t, \boldsymbol{\Theta}_1) \big)^{I(x_i \le t)} \times \prod_{i=1}^{n} \Big((1-\pi) f_2(x_i; t, T, \boldsymbol{\Theta}_2) \Big)^{I(x_i > t)},$$

where $I(x_i \le t)$ is the indicator function for the event $x_i \le t$. The corresponding complete data log-likelihood function is

$$\ell_{\text{complete}}(\boldsymbol{\Theta}; \mathcal{Y}) = \sum_{i=1}^{n} I(x_i \le t) \big(\ln \pi + \ln f_1(x_i; t^l, t, \boldsymbol{\Theta}_1) \big) + \sum_{i=1}^{n} I(x_i > t) \Big(\ln(1-\pi) + \ln f_2(x_i; t, T, \boldsymbol{\Theta}_2) \Big). \tag{6.7}$$

The complete version of the log-likelihood (6.7), as opposed to the incomplete version (6.6), is easy to optimise as it does no longer contain any CDF terms due to censored data points and allows for a separate optimisation with respect to π, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$.

However, as we do not fully observe the complete version $\mathcal{Y}$ of the data sample, the complete log-likelihood is a random variable. Therefore, it is not possible

to directly optimise the complete data log-likelihood. The intuitive idea of the EM algorithm for obtaining parameter estimates in case of incomplete data is to take the conditional expectation of the complete data log-likelihood given the incomplete data and then use this expected log-likelihood function to estimate the parameters. However, taking the expectation of the complete data log-likelihood requires the knowledge of the parameter vector, so an iterative approach is needed.

More specifically, starting from an initial guess for the parameter vector, $\boldsymbol{\Theta}^{(0)}$, the EM algorithm iterates between two steps. In the hth iteration of the E-step, we compute the conditional expectation of the complete data log-likelihood with respect to the complete data $\mathcal{Y}$ given the observed data $\mathcal{X}$ and using the current estimate of the parameter vector $\Theta^{(h-1)}$ as true values:

$$E\left(\ell_{\text{complete}}(\boldsymbol{\Theta};\mathcal{Y}) \,\middle|\, \mathcal{X};\boldsymbol{\Theta}^{(h-1)}\right).$$

In the M-step, we maximise the conditional expectation of the complete data log-likelihood obtained in the E-step with respect to the parameter vector:

$$\boldsymbol{\Theta}^{(h)} = \arg\max_{\boldsymbol{\Theta}} \; E\left(\ell_{\text{complete}}(\boldsymbol{\Theta};\mathcal{Y}) \,\middle|\, \mathcal{X};\boldsymbol{\Theta}^{(h-1)}\right).$$

Both steps are iterated until convergence.

We discuss these steps in detail for a general splicing model in the presence of random censoring in the following subsections.

6.3.3 Initial step

Before iterating the EM-steps, we need starting values for the splicing weight π and for the parameters of the distributions for the body and the tail: $\boldsymbol{\Theta}^{(0)} = (\pi^{(0)}, \boldsymbol{\Theta}_1^{(0)}, \boldsymbol{\Theta}_2^{(0)})$. Suitable starting values depend on the distributions used for the body and the tail. We discuss starting values for the splicing of the ME and Pareto distributions in Appendix D.1.1.

6.3.4 E-step

In the hth iteration of the E-step, we take the conditional expectation of the complete log-likelihood (6.7) given the incomplete data $\mathcal{X}$, the points t^l, t and T, and the current estimate $\boldsymbol{\Theta}^{(h-1)}$ for $\boldsymbol{\Theta}$. We distinguish the five cases of data points again to determine the contribution of a data point to the conditional expectation $E\left(\ell_{\text{complete}}(\boldsymbol{\Theta};\mathcal{Y}) \,\middle|\, \mathcal{X}, t^l, t, T; \boldsymbol{\Theta}^{(h-1)}\right)$:

i. $\ln \pi + E\left(\ln f_1(X_i; t^l, t, \boldsymbol{\Theta}_1) \,\middle|\, t^l \le l_i = u_i \le t < T; \boldsymbol{\Theta}_1^{(h-1)}\right)$

ii. $\ln(1-\pi) + E\left(\ln f_2(X_i; t, T, \boldsymbol{\Theta}_2) \,\middle|\, t^l < t < l_i = u_i \le T; \boldsymbol{\Theta}_2^{(h-1)}\right)$

iii. $\ln \pi + E\left(\ln f_1(X_i; t^l, t, \boldsymbol{\Theta}_1) \,\middle|\, t^l \le l_i < u_i \le t < T; \boldsymbol{\Theta}_1^{(h-1)}\right)$

iv. $\ln(1-\pi) + E\left(\ln f_2(X_i; t, T, \boldsymbol{\Theta}_2) \,\middle|\, t^l < t \le l_i < u_i \le T; \boldsymbol{\Theta}_2^{(h-1)}\right)$

v. $E\Big(\left[\ln \pi + \ln f_1(X_i; t^l, t, \boldsymbol{\Theta}_1)\right] \, I(\{X_i \le t\})$

$+ \left[\ln(1-\pi) + \ln f_2(X_i; t, T, \boldsymbol{\Theta}_2)\right] \, I(\{X_i > t\}) \,\Big|\, t^l \le l_i < t < u_i \le T; \boldsymbol{\Theta}^{(h-1)}\Big)$

Note that the event $\{t^l \le l_i = u_i \le t < T\}$ indicates that we know t^l, $l_i = u_i$, t and T, and that the ordering $t^l \le l_i = u_i \le t < T$ holds. Similar reasonings hold for the other conditional arguments in the expectations. Using the law of total expectation we can rewrite the expectation in *v.* as

$$
\begin{aligned}
&E\left(\ln \pi + \ln f_1(X_i; t^l, t, \boldsymbol{\Theta}_1) \,\middle|\, t^l \le l_i < X_i \le t < u_i \le T; \boldsymbol{\Theta}_1^{(h-1)}\right) \\
&\qquad \times P\left(X_i \le t \,\middle|\, t^l \le l_i < t < u_i \le T; \boldsymbol{\Theta}^{(h-1)}\right) \\
&+ E\left(\ln(1-\pi) + \ln f_2(X_i; t, T, \boldsymbol{\Theta}_2) \,\middle|\, t^l \le l_i \le t < X_i < u_i \le T; \boldsymbol{\Theta}_2^{(h-1)}\right) \\
&\qquad \times P\left(X_i > t \,\middle|\, t^l \le l_i < t < u_i \le T; \boldsymbol{\Theta}^{(h-1)}\right),
\end{aligned}
$$

where $\{t^l \le l_i < X_i \le t < u_i \le T\}$ denotes that t^l, l_i, t, u_i and T are known, that the ordering $t^l \le l_i < t < u_i \le T$ holds, and that $\{X_i \le t\}$. The considered conditional expectation of the complete log-likelihood is then given by

$$
\begin{aligned}
&E\left(\ell_{\text{complete}}(\boldsymbol{\Theta}; \mathcal{Y}) \,\middle|\, \mathcal{X}, t^l, t, T; \boldsymbol{\Theta}^{(h-1)}\right) \\
&= \sum_{i \in S_{i\cdot}} \left[\ln \pi + E\left(\ln f_1(X_i; t^l, t, \boldsymbol{\Theta}_1) \,\middle|\, t^l \le l_i = u_i \le t < T; \boldsymbol{\Theta}_1^{(h-1)}\right)\right] \\
&\quad + \sum_{i \in S_{ii\cdot}} \left[\ln(1-\pi) + E\left(\ln f_2(X_i; t, T, \boldsymbol{\Theta}_2) \,\middle|\, t^l < t < l_i = u_i \le T; \boldsymbol{\Theta}_2^{(h-1)}\right)\right] \\
&\quad + \sum_{i \in S_{iii\cdot}} \left[\ln \pi + E\left(\ln f_1(X_i; t^l, t, \boldsymbol{\Theta}_1) \,\middle|\, t^l \le l_i < u_i \le t < T; \boldsymbol{\Theta}_1^{(h-1)}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{i \in S_{iv.}} \left[\ln(1-\pi) + E\left(\ln f_2(X_i; t, T, \boldsymbol{\Theta}_2) \,\middle|\, t^l < t \leq l_i < u_i \leq T; \boldsymbol{\Theta}_2^{(h-1)} \right) \right] \\
&+ \sum_{i \in S_{v.}} \left[\ln \pi + E\left(\ln f_1(X_i; t^l, t, \boldsymbol{\Theta}_1) \,\middle|\, t^l \leq l_i < X_i \leq t < u_i \leq T; \boldsymbol{\Theta}_1^{(h-1)} \right) \right] \\
&\qquad \times P\left(X_i \leq t \,\middle|\, t^l \leq l_i < t < u_i \leq T; \boldsymbol{\Theta}^{(h-1)} \right) \\
&+ \sum_{i \in S_{v.}} \left[\ln(1-\pi) + E\left(\ln f_2(X_i; t, T, \boldsymbol{\Theta}_2) \,\middle|\, t^l \leq l_i \leq t < X_i < u_i \leq T; \boldsymbol{\Theta}_2^{(h-1)} \right) \right] \\
&\qquad \times P\left(X_i > t \,\middle|\, t^l \leq l_i < t < u_i \leq T; \boldsymbol{\Theta}^{(h-1)} \right).
\end{aligned} \tag{6.8}
$$

Using (6.1), the probability in the second to last term in (6.8) can be written as

$$
\begin{aligned}
&P\left(X_i \leq t \,\middle|\, t^l \leq l_i < t < u_i \leq T; \boldsymbol{\Theta}^{(h-1)} \right) \\
&\quad = \frac{F\left(t; t^l, t, T, \boldsymbol{\Theta}^{(h-1)}\right) - F\left(l_i; t^l, t, T, \boldsymbol{\Theta}^{(h-1)}\right)}{F\left(u_i; t^l, t, T, \boldsymbol{\Theta}^{(h-1)}\right) - F\left(l_i; t^l, t, T, \boldsymbol{\Theta}^{(h-1)}\right)} \\
&\quad = \frac{\pi^{(h-1)} - \pi^{(h-1)} F_1\left(l_i; t^l, t, \boldsymbol{\Theta}_1^{(h-1)}\right)}{\pi^{(h-1)} + (1-\pi^{(h-1)}) F_2\left(u_i; t, T, \boldsymbol{\Theta}_2^{(h-1)}\right) - \pi^{(h-1)} F_1\left(l_i; t^l, t, \boldsymbol{\Theta}_1^{(h-1)}\right)},
\end{aligned} \tag{6.9}
$$

and the probability in the last term of (6.8) is given by 1 minus this expression.

6.3.5 M-step

We maximise (6.8) with respect to π, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ by computing the partial derivatives and equating them to zero. In case it is not possible to find analytical solutions for one of these parameters, we need to rely on numerical procedures.

Maximisation w.r.t. π

We denote the number of observations in a set S as $\#S$ and use the notations n_1 and n_2 for the number of data points X_i smaller than or equal to t, and

above t, respectively. The partial derivative of (6.8) w.r.t. π is given by

$$\frac{\partial E\left(\ell_{\text{complete}}(\boldsymbol{\Theta};\mathcal{Y})\,\middle|\,\mathcal{X},t^l,t,T,\boldsymbol{\Theta}^{(h-1)}\right)}{\partial\pi}=\frac{n_1^{(h)}}{\pi}-\frac{n_2^{(h)}}{1-\pi}$$

with

$$n_1^{(h)}=\#S_{i\cdot}+\#S_{iii\cdot}+\sum_{i\in S_{v\cdot}}P\left(X_i\le t\,\middle|\,t^l\le l_i<t<u_i\le T;\boldsymbol{\Theta}^{(h-1)}\right),$$

and

$$n_2^{(h)}=\#S_{ii\cdot}+\#S_{iv\cdot}+\sum_{i\in S_{v\cdot}}P\left(X_i>t\,\middle|\,t^l\le l_i<t<u_i\le T;\boldsymbol{\Theta}^{(h-1)}\right).$$

Data points belonging to case v are weighted using probabilities (6.9) and $1-(6.9)$, leading to the estimates $n_1^{(h)}$ and $n_2^{(h)}$ in the hth iteration. Note that $n_1^{(h)}+n_2^{(h)}=n$. Setting the derivative equal to 0 and then solving for π yields

$$\pi^{(h)}=\frac{n_1^{(h)}}{n_1^{(h)}+n_2^{(h)}}=\frac{n_1^{(h)}}{n}. \tag{6.10}$$

This updated splicing weight can be interpreted as the proportion of data points smaller than or equal to t as estimated in the hth iteration.

Maximisation w.r.t. $\boldsymbol{\Theta}_1$

In order to maximise (6.8) w.r.t. $\boldsymbol{\Theta}_1$, we have to maximise

$$\begin{aligned}
&\sum_{i\in S_{i\cdot}}E\left(\ln f_1(X_i;t^l,t,\boldsymbol{\Theta}_1)\,\middle|\,t^l\le l_i=u_i\le t<T;\boldsymbol{\Theta}_1^{(h-1)}\right)\\
&+\sum_{i\in S_{iii\cdot}}E\left(\ln f_1(X_i;t^l,t,\boldsymbol{\Theta}_1)\,\middle|\,t^l\le l_i<u_i\le t<T;\boldsymbol{\Theta}_1^{(h-1)}\right)\\
&+\sum_{i\in S_{v\cdot}}E\left(\ln f_1(X_i;t^l,t,\boldsymbol{\Theta}_1)\,\middle|\,t^l\le l_i<X_i\le t<u_i\le T;\boldsymbol{\Theta}_1^{(h-1)}\right)\\
&\qquad\times P\left(X_i\le t\,\middle|\,t^l\le l_i<t<u_i\le T;\boldsymbol{\Theta}^{(h-1)}\right).
\end{aligned}$$

Maximisation w.r.t. $\boldsymbol{\Theta}_2$

Similarly, to maximise (6.8) w.r.t. $\boldsymbol{\Theta}_2$, we have to maximise

$$\sum_{i \in S_{ii.}} E\left(\ln f_2(X_i; t, T, \boldsymbol{\Theta}_2) \,\middle|\, t^l < t < l_i = u_i \leq T; \boldsymbol{\Theta}_2^{(h-1)}\right)$$

$$+ \sum_{i \in S_{iv.}} E\left(\ln f_2(X_i; t, T, \boldsymbol{\Theta}_2) \,\middle|\, t^l < t \leq l_i < u_i \leq T; \boldsymbol{\Theta}_2^{(h-1)}\right)$$

$$+ \sum_{i \in S_{v.}} E\left(\ln f_2(X_i; t, T, \boldsymbol{\Theta}_2) \,\middle|\, t^l \leq l_i \leq t < X_i < u_i \leq T; \boldsymbol{\Theta}_2^{(h-1)}\right)$$

$$\times P\left(X_i > t \,\middle|\, t^l \leq l_i < t < u_i \leq T; \boldsymbol{\Theta}^{(h-1)}\right).$$

6.4 Fitting the ME-Pareto model

We focus on some aspects of the specific case of the mixed Erlang – Pareto (ME-Pa) splicing model. In particular, we zoom in on how the complete data log-likelihood is constructed for the EM algorithm when a mixed Erlang distribution is used for the body, discuss how the estimation algorithm simplifies in case of no censoring and comment on the selection of splicing and truncation points.

6.4.1 Complete data log-likelihood for mixed Erlang distribution

Besides overcoming the estimation problem related to the censored data, as explained in the previous section, the EM algorithm also offers the right estimation framework when one of the splicing components is a mixture distribution. The clue is to view the data points coming from the mixture as being incomplete since the associated component-indicator vectors are not available (McLachlan and Peel, 2001). The complete data $\mathcal{Y}$ introduced above contains the uncensored sample $\boldsymbol{x} = (x_1, \ldots, x_{n_1}, x_{n_1+1}, \ldots, x_n)$ where we, without loss of generality, assume that the first n_1 data points are smaller than or equal to t. We further extend $\mathcal{Y}$ with component-indicator vectors for the first n_1 data points, denoted by $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{n_1})$ where

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ comes from the } j\text{th component density } f_E^t(\cdot; t^l, t, r_j, \theta) \\ 0 & \text{otherwise} \end{cases} \tag{6.11}$$

for $i = 1, \ldots, n_1$ and $j = 1, \ldots, M$. They are distributed according to a multinomial distribution with

$$P(\boldsymbol{Z}_i = \boldsymbol{z}_i; \boldsymbol{\beta}) = \beta_1^{z_{i1}} \ldots, \beta_M^{z_{iM}}$$

for $i = 1, \ldots, n_1$, where z_{ij} is equal to 0 or 1 and $\sum_{j=1}^{M} z_{ij} = 1$. The joint density of $(X_i, \boldsymbol{Z}_i)$ given $\{X_i \leq t\}$ equals

$$\begin{aligned} f_{X_i, \boldsymbol{Z}_i}(x_i, \boldsymbol{z}_i; t^l, t, \boldsymbol{r}, \boldsymbol{\Theta}_1) &= f_{X_i \mid \boldsymbol{Z}_i}(x_i \mid \boldsymbol{z}_i; t^l, t, \boldsymbol{r}, \boldsymbol{\Theta}_1)\, P(\boldsymbol{Z}_i = \boldsymbol{z}_i; \boldsymbol{\beta}) \\ &= \prod_{j=1}^{M} \left(f_E^t(x_i; t^l, t, r_j, \theta)\right)^{z_{ij}} \prod_{j=1}^{M} \beta_j^{z_{ij}} \\ &= \prod_{j=1}^{M} \left(\beta_j f_E^t(x_i; t^l, t, r_j, \theta)\right)^{z_{ij}} , \end{aligned}$$

for $i = 1, \ldots, n_1$. Hence the part of the complete data log-likelihood (6.7) depending on $\boldsymbol{\Theta}_1$ becomes

$$\sum_{i=1}^{n} I(x_i \leq t) \ln f_1(x_i; t^l, t, \boldsymbol{\Theta}_1) = \sum_{i=1}^{n_1} \sum_{j=1}^{M} z_{ij} \ln \left(\beta_j f_E^t(x_i; t^l, t, r_j, \theta)\right). \quad (6.12)$$

Full technical details on the EM algorithm for fitting the ME-Pareto model are treated in Appendix D.1.

6.4.2 Uncensored data

When no censoring is present, we only have data points from cases *i* and *ii*. Hence, the EM steps for π, the ME part and the Pareto part can be performed separately since the parts of the log-likelihood (6.6) containing π, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$, respectively, can then be split. We discuss this simplified setting in Appendix D.2. The splicing weight π simply gets estimated as the proportion of data points smaller than or equal to the splicing point t, see (D.16). The algorithm of Verbelen et al. (2015) is applied to fit a ME distribution to all data points smaller than or equal to t. The ξ parameter of the Pareto distribution is determined by (D.18). In case there is no upper truncation, i.e. $T = +\infty$, the solution for ξ is the Hill estimator (Hill, 1975) with threshold t. As discussed in Chapter 3, this estimator is commonly used to estimate the shape parameter ξ when modelling the tail with the Pareto distribution.

6.4.3 Selection of splicing and truncation points

Up to now, we assumed that the lower truncation point t^l, the splicing point t and the upper truncation point T are known. In many applications, there is no lower or upper truncation and we set $t^l = 0$ and $T = +\infty$.

If lower truncation is present, this boundary can often be deduced from the context. For example, in insurance, in case there is a common deductible, the lower truncation point is set to the value of this deductible.

The splicing point t might not always be as straightforward to determine. We do not propose to estimate it using a likelihood approach (see e.g. Cooray and Ananda, 2005; Lee et al., 2012). Rather, we use extreme value analysis to give an expert opinion about the choice of the splicing point. More specifically, we use the mean excess plot (Beirlant et al., 2004) to visualise where a transition from the body to the tail of the distribution is suitable. We demonstrate this type of modelling in the data examples in Section 6.6.

In situations where the upper truncation point T cannot be set based on the characteristics of the problem, as is for example the case for the earthquake magnitudes we discussed in the previous chapter, we need a strategy to decide whether upper truncation is applicable to the considered problem, and if so, an estimator for T is required. Aban et al. (2006) show that the conditional MLE for the endpoint T of a truncated Pareto distribution, if it is unknown, is given by the maximum $x_{n,n}$. The corresponding conditional MLE for ξ follows from (4.21). Beirlant et al. (2016a) further extend this methodology and provide an improved estimator for T, see (5.5). Both papers also suggest a formal test to decide between a truncated and a non-truncated tail distribution. We illustrate these methods on the first data example in Section 6.6. This approach can only be applied in case there is no censoring. For censored data, there is no method available to estimate the parameters of a truncated Pareto distribution when T is unknown.

6.5 Risk measures

In order to quantify the risk exposure of a company, several risk measures, such as the Value-at-Risk (VaR) and the Tail Value-at-Risk (TVaR), have been developed. Moreover, these risk measures can be used to determine the amount of capital to hold as a buffer against unexpected losses.

When estimating the risk measures using statistical methods, it is essential that the fitted model captures the data well. Especially a good fit of the tail part is

crucial since this corresponds to the largest losses. A global fit, hence not only a tail fit, is needed as one might be interested in computing reinsurance premiums or performing a risk analysis where focus is not only on extreme events. Further details on the estimation of risk measures can be found in McNeil et al. (2005), Klugman et al. (2012), Klugman et al. (2013) and Albrecher et al. (2017).

6.5.1 Excess-loss insurance premiums

Using a fitted splicing model such as the ME-Pa model presented here, we calculate premiums for an excess-loss insurance. For this type of insurance, the (re)insurer covers all losses above a certain retention level R. This means that she pays $(X-R)_+ = \max\{X-R, 0\}$, where X is the total claim amount. The loss for the insured (also called the cedent) is thus limited to R. This type of contract is typical in reinsurance where the reinsurer acts as the insurer's insurer and covers the losses of an insurance company above the retention level. Insurance premiums can be seen as a compensation for the (re)insurance company for bearing the risk of the insured claims, and the size of the premium is thus a measure for the risk of the claim. The net premium of such an insurance contract is given by

$$\Pi(R; t^l, t, T, \boldsymbol{\Theta}) = E((X-R)_+) = \int_R^{+\infty} (1 - F(z; t^l, t, T, \boldsymbol{\Theta}))\, dz. \tag{6.13}$$

For $t \leq R < T$ we get

$$\begin{aligned}\Pi(R; t^l, t, T, \boldsymbol{\Theta}) &= \int_R^{+\infty} \big(1 - (\pi + (1-\pi)F_2(z; t, T, \boldsymbol{\Theta}_2))\big)\, dz \\ &= (1-\pi)\Pi_2(R; t, T, \boldsymbol{\Theta}_2),\end{aligned}$$

whereas for $t^l \leq R < t$ we have

$$\begin{aligned}&\Pi(R; t^l, t, T, \boldsymbol{\Theta}) \\ &= \int_R^t \big(1 - \pi F_1(z; t^l, t, \boldsymbol{\Theta}_1)\big)\, dz + \int_t^{+\infty} \big(1 - (\pi + (1-\pi)F_2(z; t, T, \boldsymbol{\Theta}_2))\big)\, dz \\ &= (t-R) - (t-R)\pi + \pi \int_R^t (1 - F_1(z; t^l, t, \boldsymbol{\Theta}_1))\, dz + (1-\pi)\Pi_2(t; t, T, \boldsymbol{\Theta}_2) \\ &= (1-\pi)(t-R) + \pi\Pi_1(R; t^l, t, \boldsymbol{\Theta}_1) + (1-\pi)\Pi_2(t; t, T, \boldsymbol{\Theta}_2).\end{aligned}$$

Note that $\Pi(R; t^l, t, T, \boldsymbol{\Theta}) = \Pi(t^l; t^l, t, T, \boldsymbol{\Theta}) + (t^l - R)$ for $R < t^l$ and $\Pi(R; t^l, t, T, \boldsymbol{\Theta}) = 0$ for $R \geq T$.

We can rewrite

$$\begin{aligned}\Pi_1(R;t^l,t,\boldsymbol{\Theta}_1) &= \int_R^t \left(1-\frac{F_1^*(z;\boldsymbol{\Theta}_1)-F_1^*(t^l;\boldsymbol{\Theta}_1)}{F_1^*(t;\boldsymbol{\Theta}_1)-F_1^*(t^l;\boldsymbol{\Theta}_1)}\right) dz \\ &= \frac{F_1^*(t;\boldsymbol{\Theta}_1)(t-R)-(t-R)+\int_R^t(1-F_1^*(z;\boldsymbol{\Theta}_1))\,dz}{F_1^*(t;\boldsymbol{\Theta}_1)-F_1^*(t^l;\boldsymbol{\Theta}_1)} \\ &= \frac{(F_1^*(t;\boldsymbol{\Theta}_1)-1)\,(t-R)+(\Pi_1^*(R;\boldsymbol{\Theta}_1)-\Pi_1^*(t;\boldsymbol{\Theta}_1))}{F_1^*(t;\boldsymbol{\Theta}_1)-F_1^*(t^l;\boldsymbol{\Theta}_1)}\end{aligned}$$

for $t^l \leq R < t$. For the ME distribution, the premium is given by

$$\Pi_1^*(R;\boldsymbol{\alpha},\theta) = \theta^2 \sum_{m=1}^{M}\left(\sum_{l=m}^{M-1}\left(\sum_{j=l+1}^{M}\alpha_j\right)\right) f_E(R;m,\theta)$$

for $R \geq 0$, see Verbelen et al. (2015). They assume, without loss of generality, that $r_m = m$ for $m = 1,\ldots,M$. Note that $\Pi_1(R;t^l,t,\boldsymbol{\Theta}_1) = \Pi_1(t^l;t^l,t,\boldsymbol{\Theta}_1) + (t^l - R)$ for $R < t^l$ and $\Pi_1(R;t^l,t,\boldsymbol{\Theta}_1) = 0$ for $R \geq t$.

Similarly, we get

$$\Pi_2(R;t,T,\boldsymbol{\Theta}_2) = \frac{(F_2^*(T;\boldsymbol{\Theta}_2)-1)\,(T-R)+(\Pi_2^*(R;\boldsymbol{\Theta}_2)-\Pi_2^*(T;\boldsymbol{\Theta}_2))}{F_2^*(T;\boldsymbol{\Theta}_2)-F_2^*(t;\boldsymbol{\Theta}_2)}$$

for $t \leq R < T$. For the Pareto distribution we have the following premium when $R \geq t$:

$$\Pi_2^*(R;t,\xi) = \int_R^\infty \left(\frac{z}{t}\right)^{-\frac{1}{\xi}} dz = R^{-\frac{1}{\xi}+1}\frac{t^{\frac{1}{\xi}}}{\frac{1}{\xi}-1}.$$

Note that $\Pi_2(R;t,T,\boldsymbol{\Theta}_2) = \Pi_2(t;t,T,\boldsymbol{\Theta}_2) + (t-R)$ for $R < t$ and $\Pi_2(R;t,T,\boldsymbol{\Theta}_2) = 0$ for $R \geq T$.

In a pure excess-loss insurance, the potential loss for the (re)insurer is unlimited. However, the maximal amount that the (re)insurer has to pay can be limited to L. The excess-loss insurance with retention R and limit L, which is denoted as $L \,\text{xs}\, R$, has premium

$$E\big(\min\{(X-R)_+,L\}\big) = E\big((X-R)_+ - (X-(R+L))_+\big) = \Pi(R) - \Pi(R+L).$$

In practice, the limit L is typically taken as a multiple of R. More details on excess-loss insurance and limits can be found in Albrecher et al. (2017).

6.5.2 VaR, simulations and TVaR

The Value-at-Risk (VaR) is a popular risk measure and is defined as a quantile of the distribution, $\text{VaR}_{1-p} = F^{-1}(1-p)$. For the spliced distribution, the quantile function is

$$F^{-1}(p; t^l, t, T, \boldsymbol{\Theta}) = \begin{cases} F_1^{-1}(p/\pi; t^l, t, \boldsymbol{\Theta}_1) & \text{if } 0 \leq p \leq \pi \\ F_2^{-1}\big((p-\pi)/(1-\pi); t, T, \boldsymbol{\Theta}_2\big) & \text{if } \pi < p \leq 1. \end{cases}$$

The quantile function of the ME distribution F_1^{-1} cannot be computed analytically, but can be obtained by numerically inverting the CDF. For the (truncated) Pareto distribution we have

$$F_2^{-1}(p; t, T, \xi) = F_2^{*\,-1}\big(pF_2^*(T; t, \xi); t, \xi\big) = t\left(1 - p + p\left(\frac{T}{t}\right)^{-\frac{1}{\xi}}\right)^{-\xi}.$$

We can simulate losses by applying the expression for the VaR to random numbers generated from a uniform distribution on $[0, 1]$. This technique is called inverse transform sampling. Simulations are useful for aggregate loss calculations, e.g. when losses are not independent, and to determine risk measures, see Chapter 20 in Klugman et al. (2012).

Closely related is the Tail Value-at-Risk (TVaR) which is defined as the expected loss given that the loss is larger than VaR_{1-p}. It can be rewritten as (see e.g. Klugman et al., 2012)

$$\begin{aligned} \text{TVaR}_{1-p} := E\left(X \,|\, X > \text{VaR}_{1-p}\right) &= \text{VaR}_{1-p} + E\left(X - \text{VaR}_{1-p} \,|\, X > \text{VaR}_{1-p}\right) \\ &= \text{VaR}_{1-p} + \frac{E\left((X - \text{VaR}_{1-p})_+\right)}{1 - F(\text{VaR}_{1-p})} \\ &= \text{VaR}_{1-p} + \frac{\Pi(\text{VaR}_{1-p})}{p}. \end{aligned}$$

This can thus easily be computed using the expressions for VaR_{1-p} and $\Pi(R)$. Note that the last equality only holds when the CDF is continuous in VaR_{1-p} which is the case for our spliced CDF since it is continuous everywhere.

6.6 Data examples

6.6.1 Secura Re

Our first data example concerns the Secura Re dataset from Beirlant et al. (2004) which is available at `http://lstat.kuleuven.be/Wiley/Data/secura1.txt`. It consists of $n = 381$ automobile claims from Europe filed between 1988 and 2001 that are larger than 1 200 000 euro. This means that left truncation occurs at 1 200 000 euro. The claim sizes are, amongst others, corrected for inflation. Our goal is to propose a good global fit and to provide an estimate for the premium of an excess-loss insurance with a certain retention R.

The splicing point t is chosen based on the mean excess plot (Beirlant et al., 2004). This plot consists of estimates for the mean excess values

$$e(v) = E\left(X - v \,|\, X > v\right) = \frac{\int_v^{+\infty} (1 - F(x))\, dx}{1 - F(v)}, \tag{6.14}$$

in the order statistics $v = X_{n-k,n} = \hat{Q}\left(1 - \frac{k+1}{n+1}\right) = \hat{Q}\left(\frac{n-k}{n+1}\right)$ with $k = 1, \ldots, n-1$, where the CDF F is estimated by the empirical CDF $\hat{F}$, and $\hat{Q}$ is the corresponding empirical quantile function. The horizontal part on the left in the mean excess plot in Figure 6.2 indicates that a distribution with an exponential-like tail is suitable there, whereas the linear increasing part suggests a Pareto tail. On the right, there is a decreasing trend which suggests that there might be an upper truncation point. The splicing point is chosen at the transition of the horizontal part to the linear increasing part as indicated by the vertical dashed line. This point $t = 2\,600\,000$ lies very close to the value 2 580 026 that is determined in Beirlant et al. (2004) using adaptive threshold selection methods.

We fit the ME-Pareto splicing model starting from $M = 10$, and consider spread factors $s \in \{1, \ldots, 10\}$ (see Appendix D.1.1). The fitted model was obtained using $s = 1$ and is summarised in Table 6.1. It consists of a single Erlang distribution for the body and the Pareto distribution for the tail. Based on the mean excess plot, a splicing model with an upper truncated Pareto distribution provides an alternative possible tail model. Fitting a ME and truncated Pareto splicing model, as discussed in Section 6.4.3, with the same splicing point then gives $\hat{\xi} = 0.298$ and $\hat{T} = 9\,387\,484$ whereas the other parameters remain the same.

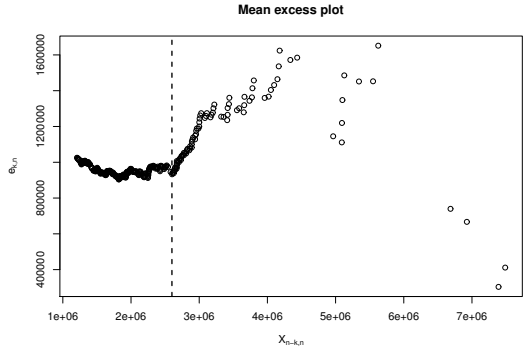


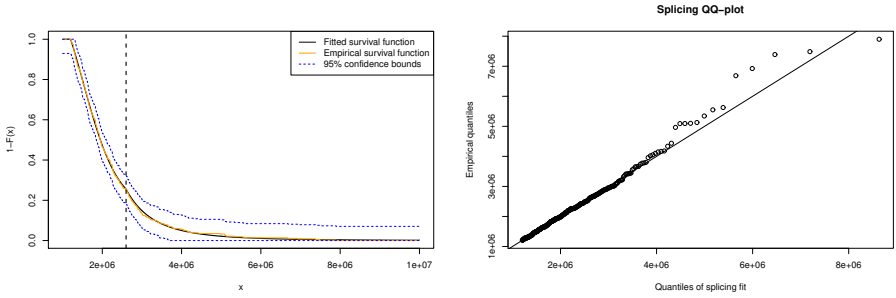
Figure 6.2: Secura Re: Mean excess plot.

Splicing	ME	Pareto
$\hat{\pi} = 0.744$	$\hat{\alpha} = 1$	$\hat{\xi} = 0.263$
$t^l = 1\,200\,000$	$\hat{r} = 8$	
$t = 2\,600\,000$	$\hat{\theta} = 217\,084$	
$T = +\infty$		

Table 6.1: Secura Re: summary of the fitted ME-Pa splicing model.

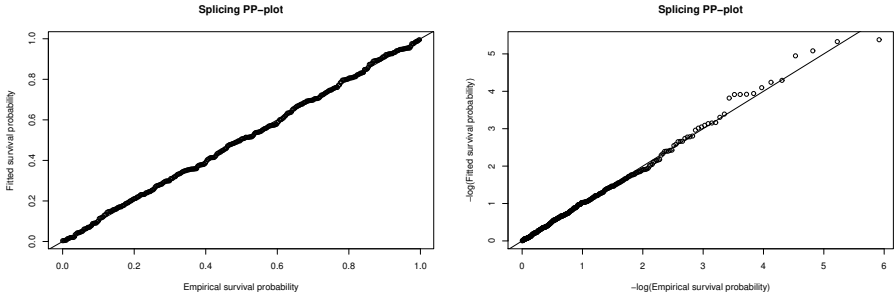
In order to evaluate the splicing fit with the ME and Pareto distributions, graphical tools, information criteria and goodness-of-fit (GoF) tests are considered. A first graphical tool is the survival plot in Figure 6.3a where the fitted survival function (dark) is plotted together with the empirical survival function (light). 95% confidence bands for the empirical estimator (dashed) and a vertical line indicating the splicing point are also added. These confidence bands are determined using the Dvoretzky-Kiefer-Wolfowitz inequality (Massart, 1990). The fitted spliced survival function follows the empirical survival function closely and lies well within the confidence bands. Next, to inspect this fit in more detail, a QQ-plot is constructed (Figure 6.3b) comparing the empirical quantiles to the fitted quantiles. The points on the QQ-plot are close to the 45 degree line suggesting a good fit. Closely related is the probability-probability (PP) plot in Figure 6.4a where the fitted survival function is plotted vs. the empirical survival function. This plot confirms that the model gives a good global fit. However, it is difficult to asses the quality of the tail fit from the PP-plot. Therefore, a PP-plot with a minus-log scale is also constructed (Figure 6.4b). The upper

right corner then corresponds to the tail of the distribution. As expected, there are some deviations from the 45 degree line for the largest points, but the plot still indicates a good global fit.



(a) (b)

Figure 6.3: Secura Re: (a) Survival plot and (b) QQ-plot of the fitted ME-Pa splicing model.



(a) (b)

Figure 6.4: Secura Re: PP-plots of the fitted ME-Pa splicing model with (a) ordinary and (b) minus-log scale.

Additional to the graphical tools, we look at the negative log-likelihood (NLL), AIC and BIC values for each model where lower values are better, see Table 6.2. The AIC and BIC are defined as

$$\text{AIC} = 2 \times \text{NLL} + 2 \times df \quad \text{and} \quad \text{BIC} = 2 \times \text{NLL} + \ln n \times df$$

where df denotes the degrees of freedom, i.e. the number of estimated parameters in the model. Moreover, we consider the Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) GoF statistics as they are a measure for the distance between the empirical CDF and the fitted CDF of a model. The KS statistic is

defined as

$$D_n = \sup_{x \geq t^l} |\hat{F}_n(x) - F(x)|$$

where $\hat{F}_n$ is the empirical CDF based on n observations and F the fitted CDF. The AD statistic is given by

$$A_n = n \int_{t^l}^{+\infty} \frac{(\hat{F}_n(x) - F(x))^2}{F(x)(1 - F(x))}\, dx.$$

Note that both test statistics take lower truncation at t^l into account. These statistics are commonly used to test if the data sample is drawn from a specified (continuous) distribution. The standard P-values of the test are not valid when the model parameters are estimated from the data (Babu and Rao, 2004). Therefore, we use a bootstrap approach that is detailed in Babu and Rao (2004) and Klugman et al. (2012). First, we compute the KS and AD test statistics using the fitted model for the Secura data. Then, we generate 1000 samples with replacement from the Secura data. For each sample, the model is fitted and then the KS and AD statistics are computed. The P-values are then obtained as the proportion of these 1000 test statistics that exceed the test statistic computed in the first step. The `R` (`R` Core Team, 2017) packages *stats* (KS) and *ADGofTest* (Gil Bellosta, 2011) (AD) are used to compute the test statistics. The results are also displayed in Table 6.2 where values closer to 0 indicate a better fit. The corresponding P-values are added between brackets. Apart from the fitted splicing model, we also consider the following models:

- The ME and truncated Pareto splicing model (ME-TPa) as discussed above.
- The ME and GPD splicing model (ME-GPD) with the same splicing point as before: $t = 2\,600\,000$. Hence, it has the same ME distribution for the body of the distribution as the ME-Pa and ME-TPa splicing models. The tail of the distribution is modelled by a GPD with parameters $\hat{\xi} = 0.3512$ (shape) and $\hat{\sigma} = 626\,554.8$ (scale).
- The splicing model of Beirlant et al. (2004) combining the exponential distribution and the Pareto distribution (Exp-Pa) with the same splicing point as before: $t = 2\,600\,000$. The ML estimates for the parameters are $\hat{\lambda} = 1/1\,397\,147$ (rate of the exponential distribution) and $\hat{\xi} = 0.263$ (shape of the Pareto distribution).
- The mixed Erlang fit of Verbelen et al. (2015): $t^l = 1\,200\,000$, $\hat{\boldsymbol{\alpha}} = (0.971, 0.029)$, $\hat{\boldsymbol{r}} = (5, 16)$ and $\theta = 360\,096$.

Note that the first three models are fitted using our general fitting procedure, and the ME fit is obtained using the approach in Verbelen et al. (2015).

The simpler exponential-Pareto model (Exp-Pa) provides a worse fit than the ME-Pa, ME-GPD and ME models since the NLL, AIC and BIC values are higher. The splicing model with the truncated Pareto distribution (ME-TPa) is performing slightly worse than the untruncated model (ME-Pa). The test for truncation of a Pareto tail (Beirlant et al., 2016a) gives a P-value of 0.3889, at the splicing point t, which means that the null hypothesis of a non-truncated Pareto tail is not rejected on the 5% significance level. This confirms that the untruncated Pareto distribution might be more suitable. However, for every case one should decide what type of tail behaviour is appropriate and whether a bounded model has any economic or physical meaning. Based on the NLL, AIC and BIC values, we can conclude that the ME-Pa fit slightly improves the ME fit although differences are small. The values of the NLL suggest that the ME-GPD fit is slightly better than the ME-Pa fit. However, when taking both the quality of the fit and the number of parameters into account, as is done in the AIC and BIC, the ME-Pa model is preferred over the ME-GPD model. The P-values of the GoF tests are large for all models suggesting that all models provide an appropriate fit for the data. Based on the graphical tools, the ICs, and the P-values of the KS and AD tests, we propose to use the ME-Pa model when modelling the Secura Re data.

Model	NLL	AIC	BIC	KS	AD
Exp-Pa	5502.27	11 010.53	11 022.28	0.0364 (0.875)	0.6853 (0.740)
ME-Pa	5499.13	**11 006.26**	**11 021.93**	0.0221 (0.997)	0.2173 (0.987)
ME-TPa	**5498.54**	11 007.07	11 026.65	0.0280 (0.967)	0.2764 (0.974)
ME-GPD	5498.96	11 007.91	11 027.49	0.0221 (0.995)	0.1919 (0.995)
ME	5499.99	11 007.99	11 023.65	0.0237 (0.995)	0.1889 (0.999)

Table 6.2: Secura Re: NLL, AIC and BIC values, and GoF test statistics and P-values.

As an illustration, premiums for excess-loss insurances can be computed using the fitted models. Table 6.3 shows the computed premiums for different models and different retentions. Additional to the five previously mentioned models, premiums are also computed non-parametrically using (6.13) with the empirical survival function (Non-par.), and using the combination of a non-parametric fit for the body (below $t = 2\,600\,000$ as before) and the Pareto distribution for the tail (Non-par.–Pa). All five parametric models result in premiums that are close to the ones obtained using the non-parametric model when the retention levels are small. For higher levels the estimates are substantially different. The non-parametric model results in zero premiums when the retention levels are larger than the maximal data value, 7 898 639. Similarly, premium estimates are 0 for insurances with retention levels that are larger than $\hat{T}$ when the ME-TPa

splicing model is used. The ME distribution, which has an exponential tail, results unsurprisingly in lower premium estimates for high retentions than the heavy-tailed ME-Pa and ME-GPD models. The Exp-Pa and ME-Pa models have the same fit for the tail, but a different fit for the body. Therefore, the premium estimates for high retentions are the same, but the premiums for retentions below the splicing point $t = 2\,600\,000$ differ. Although the ME-Pa, ME-TPa and ME-GPD models have the same model for the body of the distribution, the estimates for the premiums also differ for low retentions since the survival function is integrated starting from the retention level when estimating the premiums, see (6.13).

R	Non-par.	Non-par.–Pa	Exp-Pa	ME-Pa	ME-TPa	ME-GPD	ME
1 200 000	1 030 667	1 031 738	1 031 738	1 031 738	1 042 430	1 040 995	1 030 667
2 000 000	445 330	446 401	456 172	447 054	457 746	456 311	444 751
3 000 000	161 728	159 527	159 527	159 527	168 159	170 187	164 585
4 000 000	74 696	71 350	71 350	71 350	71 036	84 821	78 228
5 000 000	35 888	38 225	38 225	38 225	32 106	51 215	39 696
7 500 000	1075	12 298	12 298	12 298	3103	21 558	4025
10 000 000	0	5501	5501	5501	0	11 987	159

Table 6.3: Secura Re: estimates for premiums of excess-loss insurance with different retentions R.

6.6.2 Motor third party liability insurance

The second data example consists of motor third party liability (MTPL) insurance claims in Europe between 1995 and 2010 (Albrecher et al., 2017). They are evaluated at the end of 2010, i.e. right before the beginning of 2011, and 59% of the 837 claims are not closed at that time. All amounts are indexed in order to reflect costs in calendar year 2011, with inflation taken into account. Our goal is again to provide a good overall fit and to estimate excess-loss insurance premiums.

As discussed in Section 6.1, a significant time may elapse between the claim occurrence and its final settlement (due to e.g. legal procedures or severe bodily injury). In order to illustrate the development of a claim, in Figure 6.5, we show for four claims the cumulative indexed payment (full line) and the indexed incurred (dashed line) at the end of each year. The incurred at the end of a given year is equal to the sum of the cumulative payment up to that moment and an expert's estimate for the outstanding loss. The claims occurred, respectively, in 1995, 1996, 1997 and 1998. The first and third claim are closed before the end of the observation period (indicated by the vertical dashed line in Figure 6.5), and hence the cumulative indexed payment and the indexed incurred value at the

end of 2010 are equal. The second and fourth claim are still in development at the end of 2010 and the indexed incurred is larger than the cumulative indexed payment at that moment.

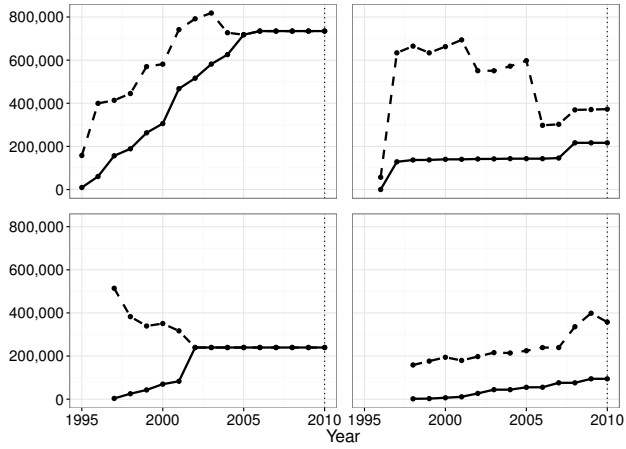


Figure 6.5: MTPL: cumulative indexed payments (full line) and indexed incurred values (dashed line) at the end of each year for four claims. The moment of evaluation, i.e. the end of 2010, is indicated by the vertical dashed line.

We apply the splicing approach for censored data using an interval censoring framework with the cumulative indexed payments at the end of 2010 as lower bound for the final cumulative indexed payment. It makes sense to construct an upper bound based on the incurreds since they are determined conservatively using information on the specific claim: e.g. the severity of the accident, the number of people involved. As an illustration of the method, and by lack of further claim information, we use here the indexed incurreds at the end of 2010 as upper bound. However, when a claim is early in development, i.e. there is a small period between the claim occurrence and the moment of evaluation, the incurreds might still be too uncertain to be used as an upper bound since the information available to the expert might be limited. After several years of development, the quality of the incurreds has improved a lot, as more information becomes available, making them more suitable as an upper bound. Since claims with accident years between 2006 and 2010 are still early in development, and we do not have more information to improve their incurreds, we omit them for the analysis (as is done in Albrecher et al., 2017). We then have 596 claims left and 45% of them are not closed at the end of 2010. A more prudent approach is to only use the cumulative indexed payments at the end of 2010 as lower bound in a right censoring framework. However, this does not take the valuable information of incurreds into account. Another possibility is

to ignore any censoring information and to only consider the indexed incurreds when estimating the final claim amount. In this example we compare these three possible strategies.

As before, we rely on the mean excess plot to choose the splicing point t. We now use the Turnbull estimator (Turnbull, 1976) to estimate the distribution function in (6.14). This is a non-parametric estimator for the CDF in the case of interval censored data points. It extends the Kaplan-Meier estimator (Kaplan and Meier, 1958), which can only be used for right censored data, to interval censored data. There is no analytical solution for the Turnbull estimator and its computation relies on the EM algorithm. We use the implementation in the `R` package *interval* (Fay and Shaw, 2010). The resulting mean excess estimates are

$$\hat{e}(v) = \frac{\int_v^{+\infty}(1 - \hat{F}^{TB}(x))\,dx}{1 - \hat{F}^{TB}(v)}, \tag{6.15}$$

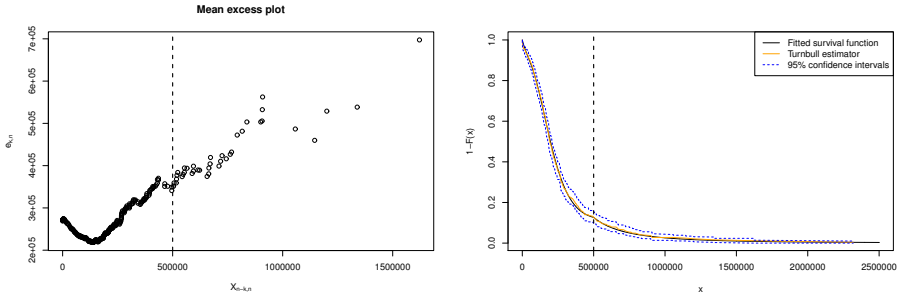
where $\hat{F}^{TB}$ is the Turnbull estimator for the CDF. We evaluate this function in $v = \hat{Q}^{TB}(1-(k+1)/(n+1)) = \hat{Q}^{TB}((n-k)/(n+1))$, for $k = 1, \ldots, n-1$, where $\hat{Q}^{TB}$ is the estimator for the quantile function based on the Turnbull estimator, since in the uncensored case we also used the empirical quantiles corresponding to $1/(n+1), \ldots, (n-1)/(n+1)$. The estimates are plotted in Figure 6.6a. The mean excess plot now has a convex shape indicating that a Pareto tail is suitable. A different slope is visible after 500 000 and we therefore choose the splicing point at $t = 500\,000$ as shown by the vertical line. As discussed in Section 6.3.1, there are five classes of data points when fitting a splicing model to censored data. Using the splicing point $t = 500\,000$, the number of data points per class is $\#S_{i.} = 296$, $\#S_{ii.} = 34$, $\#S_{iii.} = 175$, $\#S_{iv.} = 25$ and $\#S_{v.} = 66$.

The model is fitted starting from $M = 10$ and with $s \in \{1, \ldots, 10\}$ (see Appendix D.1.1). The fitted model consists of $M = 2$ Erlangs and was obtained using $s = 2$. It is summarised in Table 6.4.

Splicing	ME	Pareto
$\hat{\pi} = 0.873$	$\hat{\boldsymbol{\alpha}} = (0.171, 0.829)$	$\hat{\xi} = 0.438$
$t^l = 0$	$\hat{\boldsymbol{r}} = (1, 4)$	
$t = 500\,000$	$\hat{\theta} = 55\,227$	
$T = +\infty$		

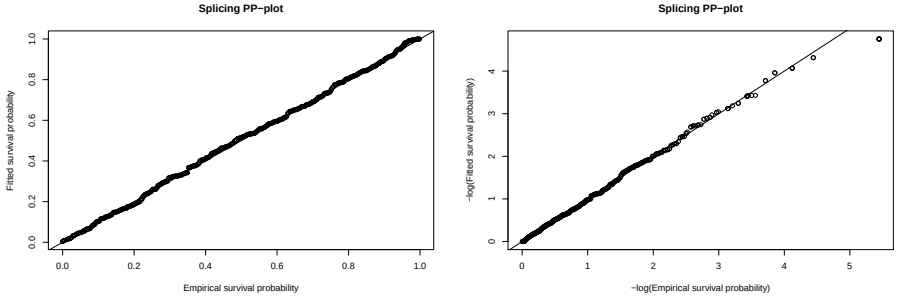
Table 6.4: MTPL: summary of the fitted ME-Pa splicing model.

Some of the graphical tools used in Section 6.6.1 can be extended to the censoring case. The fitted survival function can be compared to the non-parametric Turnbull estimate (Figure 6.6b). Pointwise confidence intervals are obtained using 200 bootstrap samples generated by the `R` package *interval* (Fay and Shaw, 2010). They are added as dashed lines in Figure 6.6b. The fitted survival function follows the Turnbull estimate closely and stays within the confidence intervals suggesting a good fit. PP-plots are made using the fitted survival function and the Turnbull survival function, see Figures 6.7a and 6.7b, where a minus-log scale is used in the second plot. Both lines are close to the 45 degree line indicating that the fitted model is suitable for the data.



(a) (b)

Figure 6.6: MTPL: (a) Mean excess plot based on the Turnbull estimator and (b) survival plot of the fitted ME-Pa splicing model.



(a) (b)

Figure 6.7: MTPL: PP-plots of the fitted ME-Pa splicing model with (a) ordinary and (b) minus-log scale.

Figure 6.8 shows estimates for premiums of excess-loss insurance for different retentions. The premiums are estimated using the considered splicing model in the interval censoring framework (full line), and compared to estimates obtained using a splicing model based on the right censoring framework (dashed line) and a splicing model without censoring using the indexed incurreds (dash-dot line). The second approach gives higher premium estimates than the first approach since the (censored) total amount paid for each claim is not bounded from above. The incurreds are conservative expert estimates of the final cumulative claim amount. Only using the indexed incurreds in an uncensored framework does not take into account that the actual total amount that needs to be paid can be lower than the indexed incurreds. Therefore, it leads to higher premium estimates than for the splicing model using interval censored data. Using all information available, the cumulative indexed payments and the indexed incurreds, leads to significantly lower premium estimates.

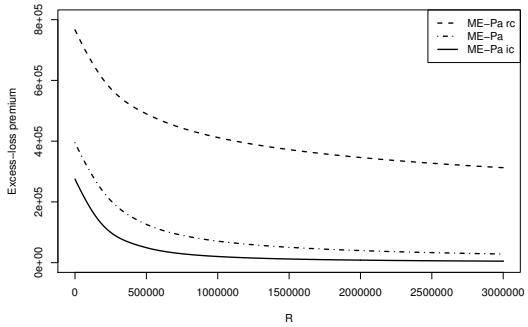


Figure 6.8: MTPL: Estimates for premiums of excess-loss insurance with different retentions using ME-Pa splicing model with interval censoring (full line), right censoring (dashed line) and no censoring (dash-dot line).

As mentioned in the introduction of this chapter, a non-parametric fit for the body and a parametric model (e.g. Pareto distribution) for large losses can be used instead of a splicing model. When censoring is present, this approach can no longer be used as we might have data points of class v (see Figure 6.1) where the lower bound of the interval is in the body of the distribution, whereas the upper bound is in the tail. As is shown in this example, our general splicing framework can handle observations of this type and can hence be used to provide a global fit. This global fit is e.g. needed to compute premiums for excess-loss insurances.

We discussed the possibility to use the GPD instead of the Pareto distribution in the splicing model. Unlike for the Pareto distribution, the fourth and sixth

expectation in the E-step (6.8) cannot be computed analytically when using the GPD (with $\xi \neq 0$). This makes the whole procedure numerically more intensive as it requires numerical integration. Without censoring, this drawback is not present as only the first two expectations in the E-step (6.8) need to be computed.

6.7 Conclusions

In order to get a suitable global fit for financial loss data we propose a new splicing model. It combines the flexibility of the mixed Erlang distribution to model the body of the distribution with the Pareto distribution to provide a suitable fit for the tail. Hence, our proposal avoids ad hoc combinations of a standard light-tailed distribution for the body with a heavy-tailed distribution for the tail.

Motivated by real life insurance datasets where censoring and truncation are omnipresent, we provide a general framework for fitting a spliced distribution to censored and/or truncated data. This fitting procedure uses the EM algorithm to handle data incompleteness due to censoring. Moreover, we give details on the application of this procedure to fit the ME-Pa model.

Estimates for excess-loss premiums and risk measures such as the VaR can be easily extended to the splicing context. We illustrate the flexibility of the proposed ME-Pa splicing approach using the lower truncated Secura Re dataset and using the MTPL dataset where censoring is present.

As we provide a general procedure to fit a splicing model to censored and/or truncated data, other distributions for the body and/or tail can be considered. We illustrate the use of the GPD instead of the Pareto distribution for the tail.

Chapter 7

Conclusions and further research perspectives

7.1 Conclusions

In Chapter 3, we investigated, based on EVT, if the recent financial crisis was a Black Swan event. We looked at two indicators: 1. the return periods for the experienced losses in view of the pre-crisis data, and 2. tests for significant differences in the scale or shape parameters of the Pareto tail before and after the crisis. We developed new estimators for the scale parameter, and provided asymptotic results for weakly-dependent data. We argued that Barclays can be considered as having experienced a Black Swan event whereas this is not the case for Credit Suisse. Based on economic indicators of both banks, we concluded that Barclays was indeed more vulnerable than Credit Suisse.

Motivated by earthquake magnitude data and river flow data, we extended the approach from Aban et al. (2006) and Beirlant et al. (2016a) to truncated distributions whose parent distributions have EVI $\xi > -1/2$. Simulations and data examples in Chapter 4 and Chapter 5 illustrated that the new estimator based on the POT approach works for both truncated heavy tails and truncated light tails.

In Chapter 5, we used this new approach to estimate the maximum possible earthquake magnitude in Groningen where earthquakes are induced by gas extraction. Moreover, we looked at upper confidence bounds to quantify the uncertainty in the estimation of this endpoint. Using the different considered

techniques, we find estimates for the maximum possible earthquake magnitude in Groningen between 3.65 and 3.9 on the Richter scale. 90% upper confidence bounds based on these methods range from 4 to 4.65. Based on simulations, we can conclude that this estimator, and the estimator of Beirlant et al. (2016a) applied to the earthquake energy, are a valuable addition to the existing methods for estimating the maximum possible earthquake magnitude. Moreover, our EVA suggests that an upper truncated exponential distribution, and hence the Gutenberg-Richter distribution, is indeed a suitable model for the earthquake magnitudes in Groningen.

In the last chapter, we proposed a global fit for loss data using a splicing model with the flexible ME distribution for the body, and the Pareto distribution or GPD for the tail. This avoids ad hoc combinations of a standard light-tailed distribution for the body with a heavy-tailed distribution for the tail. Moreover, we provided a procedure to fit a general splicing model to censored and/or truncated data using the EM algorithm. Using two (re)insurance examples, the lower truncated Secura Re data and the MTPL data where censoring is present, we illustrated the flexibility of the ME-Pa splicing approach.

There are several widely used `R` packages related to EVT: *actuar* (Dutang et al., 2008), *evir* (Pfaff and McNeil, 2012), *fExtremes* (Würtz and Rmetrics Association, 2013) and *QRM* (Pfaff and McNeil, 2016) which accompanies McNeil et al. (2005). An overview of the (main) `R` packages related to EVT can be found in the CRAN task view "Extreme Value Analysis" (Dutang and Jaunatre, 2017). These packages contain implementations of extreme value distributions, classical extreme value plots, estimators for the EVI and the POT approach. However, several EVT estimators and plots, especially those adapted for censoring or truncation, were not available. We implemented many of these methods in the *ReIns* package (Reynkens and Verbelen, 2017) which complements Albrecher et al. (2017). This provides a unified framework for all estimators and plots. The *ReIns* package contains:

- Basic EVT estimators and graphical methods as described in Beirlant et al. (2004) and Albrecher et al. (2017): QQ-plots, the Hill estimator, the POT approach, the scale estimators proposed in Chapter 3, etc.
- Several extreme value distributions such as the Pareto distribution, the Burr distribution and the GPD. Moreover, upper truncated distributions such as the truncated Pareto distribution and the truncated GPD are also included.
- EVT estimators and graphical methods adapted for censored or truncated data as described in Beirlant et al. (2007) and Einmahl et al. (2008), and Aban et al. (2006), Beirlant et al. (2016a) and Chapter 4, respectively.

- Splicing of the mixed Erlang distribution with EVT distributions (Pareto, GPD), including the procedure for fitting the ME-Pa splicing model to interval censored data, as introduced in Chapter 6.
- Risk measures as described in Chapter 6: Value-at-Risk, Tail Value-at-Risk and excess-loss premium estimates.

The package is available on CRAN: `https://CRAN.R-project.org/package=ReIns` where an introduction can also be found.

7.2 Further research perspectives

As indicated in Chapter 5, bias reduction of the truncated EVT estimators of Aban et al. (2006) and Chapter 4 is needed. A possible solution for bias reduction of the truncated Pareto estimator is to consider the upper truncated EPD, and extend the likelihood approach of Beirlant et al. (2009) with ideas from Aban et al. (2006) and Beirlant et al. (2016a).

In order to reduce the effect of the earthquakes, the Dutch government lowered the production from 54 billion cubic metres in 2013 to 24 billion cubic metres in 2016 (van den Beukel, 2016). It is important to quantify the effect of these production measures on the seismicity. Another influential effect is that the seismic moment per unit gas produced increases when the reservoir gets emptier (Bourne et al., 2014). The maximum possible earthquake magnitude is not time-dependent as mentioned earlier, but the maximum expected earthquake magnitude does depend on the production regime and the activity rate. Zöller and Holschneider (2016b) also provide estimates for this quantity for the Groningen case. It would be interesting to take time-dependence into account in our models and propose suitable estimators for the maximum expected earthquake magnitude.

As shown in Beirlant et al. (2004), the mean excess function of the logarithm of the data,

$$E\left(\ln X - \ln v \,|\, X > v\right) = \frac{\int_v^{+\infty} (1 - F(x)) \, \frac{dx}{x}}{1 - F(v)}, \tag{7.1}$$

converges for Pareto-type distributions to ξ as $v \to \infty$. They also note that the Hill estimator $H_{k,n}$ can be obtained from (7.1), for $v = X_{n-k,n}$, by estimating the CDF F by the empirical CDF $\hat{F}$. In the same manner, Worms and Worms (2014) propose to estimate the EVI under right censoring by estimating the CDF using the Kaplan-Meier estimator. To extend the Hill estimator to interval

censored data, we propose to estimate the CDF by the Turnbull estimator $\hat{F}^{TB}$, for $v = \hat{Q}^{TB}(1 - (k+1)/(n+1))$ as in (6.15). Further research is needed to investigate the asymptotic and finite sample behaviour of this estimator.

For the MTPL data example, the random censoring assumption might not be satisfied since claims with longer development times appear to be more likely to be censored, see Section 4.4.3 in Albrecher et al. (2017). However, their analysis indicates that this assumption does hold conditional on a certain development time. Moreover, we expect that claims with longer development times have heavier tails since large claims take on average longer to be closed. Therefore, it might be useful to consider a regression approach for Pareto-type tails with the development time as covariate. However, the development time is also right censored which complicates the problem. Akritas and Van Keilegom (2003) propose an estimator for the conditional CDF where both the response and covariate can be right censored. A first approach would be to estimate the CDF in (7.1) by their estimator which gives a conditional estimator for the EVI under right censoring. As this conditional estimator uses kernel-based weights, a proper selection criterion for the bandwidth is needed which requires investigation of the asymptotic properties of the estimator. Another interesting idea is to consider a splicing model conditional on the development time. The fitting procedure of Chapter 6 can then for example be extended by introducing the kernel-based weights in the likelihood.

In practice, reinsurance forms are often combined across various lines of business (LOBs). Then, not only the different LOBs, but also the dependence between them needs to be modelled. An example of a multivariate dataset is the Danish fire insurance dataset (Rytgaard, 1996) from the Copenhagen Reinsurance Company which contains information on 2167 fire losses from 1980 to 1990. For each claim, the total loss is divided into damage to building ($X_{i,1}$), damage to content ($X_{i,2}$) and loss of profits ($X_{i,3}$) where all variables are expressed in millions of Danish Krone. A claim is only considered if the total loss exceeds 1 million Danish Krone, i.e. $X_{i,1} + X_{i,2} + X_{i,3} \geq 1$. Scatter plots of the log-transformed data are shown in Figure 7.1. This dataset has been considered in many books and papers, see e.g. McNeil (1997), Embrechts et al. (1997), Drees and Müller (2008) and Albrecher et al. (2017). Another multivariate example is given in Frees and Valdez (1998). They jointly model losses and allocated loss adjustment expenses using copulas.

The ME distribution with common scale parameter can also be extended to higher dimensions (Lee and Lin, 2012). They prove that the class of MME distributions is dense in the space of positive continuous multivariate distributions in the sense of weak convergence, extending the result from Tijms (1994) for the class of univariate ME distributions. Willmot and Woo (2015)

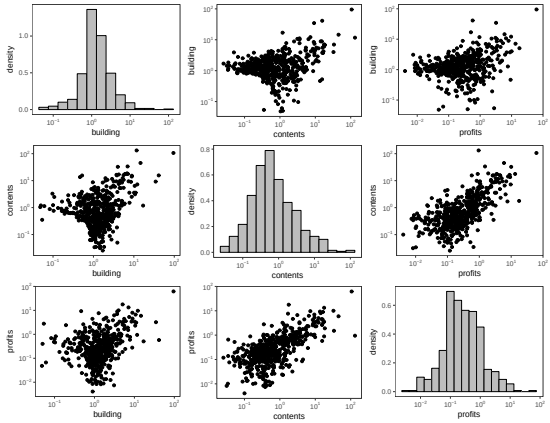


Figure 7.1: Danish fire insurance data: scatterplot matrix on log-scale.

study the analytical properties of the MME class and motivate their use in actuarial science. Verbelen et al. (2016) extend the fitting methodology of Lee and Lin (2012) to take censoring and truncation into account.

The foundations of EVT have also been extended to higher dimensions, see Chapter 8 in Beirlant et al. (2004) for an overview of multivariate MDAs and multivariate extreme value distributions. Recently, most attention has been paid to extending the POT approach to higher dimensions where the excesses over a high threshold are modelled using the multivariate generalised Pareto distribution (MGPD), see e.g. Rootzén and Tajvidi (2006), Falk and Guillou (2008), Rootzén et al. (2016) and Kiriliouk et al. (2016). Note that in this multivariate setting, a point is extreme if at least one of its components exceeds this threshold. The MGPD can be obtained by combining univariate GPDs using a certain dependence structure, e.g. a symmetric logistic model.

Verbelen et al. (2016) note that the MME has the same problem with heavy-tailed data as the univariate ME. A possible solution to provide a global fit for multivariate heavy-tailed, dependent data is to combine the MME distribution and the MGPD in a multivariate splicing model. Figure 7.2 contains the CDF of the bivariate MME-MGPD splicing model for "building" and "contents" fitted to the reduced Danish fire insurance data where each component exceeds 1 million. The MME distribution models bivariate losses below the splicing point $(7.32, 10.27)$ (indicated by the black dots) and the MGPD fits losses that are larger than this splicing point in at least one dimension (white dots). More details on the fitting procedure can be found in Section 4.5 in Albrecher et al. (2017).

Further research is needed regarding the choice of the splicing point in higher dimensions and estimation of the dependence structure for the MGPD. Moreover, it would be interesting to extend the fitting procedure to take censoring into account.

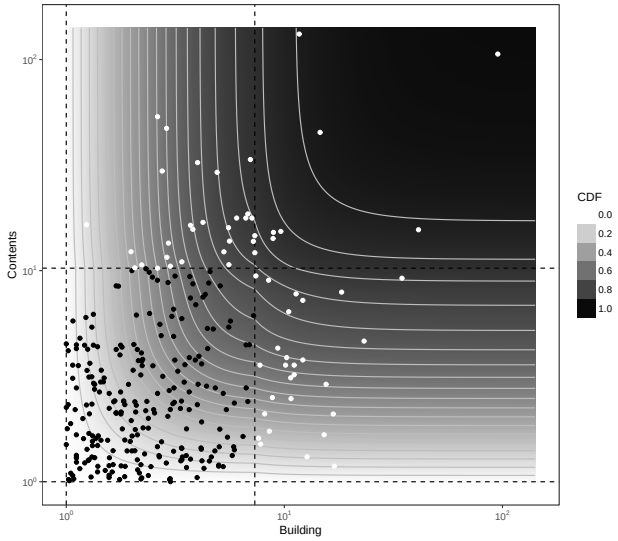


Figure 7.2: Danish fire insurance data: CDF of the fitted bivariate MME-MGPD splicing model.

Appendix A

Appendix for Chapter 3

A.1 Derivation of the scale estimators $\hat{A}_{k,n}$ and $\hat{A}^{EP}_{k,n}$

Starting from the Hall model (3.7) and ignoring the second order terms yields the approximation

$$\bar{F}(x) \sim Ax^{-1/\xi}, \text{ as } x \to \infty. \tag{A.1}$$

Alternatively, for intermediate order statistics $X_{n-k,n}$, the tail probability $\bar{F}(X_{n-k,n})$ can be estimated by the empirical probability $k/n \approx (k+1)/(n+1)$, leading to the defining equation

$$\hat{A}_{k,n} X_{n-k,n}^{-1/H_{k,n}} = \frac{k+1}{n+1},$$

where ξ in (A.1) is estimated by the Hill estimator $H_{k,n}$. This immediately gives (3.12).

In order to reduce the bias in estimating the scale parameter, ξ first needs to be estimated by the EPD estimator $\hat{\xi}_{k,n}$ to lift up the bias caused by the estimation of ξ. The other source of bias originates from ignoring the second order terms when approximating A. Following a similar reasoning as before, now taking the second order terms into account, the defining equation is

$$\hat{A} X_{n-k,n}^{-1/\hat{\xi}_{k,n}} \left(1 + bX_{n-k,n}^{-\beta}(1+o(1))\right) = \frac{k+1}{n+1}.$$

Since $\kappa = \kappa_t = \xi b t^{-\beta}(1+o(1))$, we can estimate $bX_{n-k,n}^{-\beta}(1+o(1))$ by $\hat{\kappa}_{k,n}/\hat{\xi}_{k,n}$ with $\hat{\kappa}_{k,n}$ the EPD estimator for κ at the threshold $t = X_{n-k,n}$. In order to

obtain numerically stable results, we can use that $(1+\kappa_t/\xi)^{-1} \sim 1-\kappa_t/\xi$ since $\kappa_t \to 0$ as $t \to \infty$, which leads to the bias reduced scale estimator in (3.13).

A.2 Proofs for Section 3.3

Proof of Theorem 3.1. Remark that

$$\sqrt{k}\left(\ln \hat{A}_{k,n} - \ln A\right) = T_{k,n}^{(1)} + T_{k,n}^{(2)}$$

with

$$T_{k,n}^{(1)} = \sqrt{k}\left(\frac{\ln X_{n-k,n}}{H_{k,n}} - \frac{\ln U(n/k)}{\xi}\right)$$

$$T_{k,n}^{(2)} = \sqrt{k}\left(\frac{\ln U(n/k)}{\xi} + \ln\left(\frac{k+1}{n+1}\right) - \ln A\right).$$

First, as $U(x) = A^{\xi}x^{\xi}\left(1+\xi b A^{-\xi\beta}x^{-\xi\beta}(1+o(1)\right)$ when $x \to \infty$,

$$T_{k,n}^{(2)} = -\sqrt{k}B(n/k)\frac{1}{\xi\beta}(1+o(n/k)),$$

as $n/k \to \infty$. Next, with $\tilde{H}_{k,n} := \frac{1}{k}\sum_{j=1}^{k}\left(\ln X_{n-j+1,n} - \ln U(n/k)\right)$ and $E(\tilde{H}_{k,n}) = \xi + B(n/k)/(1+\xi\beta)$, see Hsing (1991), we get

$$\begin{aligned}
T_{k,n}^{(1)} &= -\frac{\ln U(n/k)}{H_{k,n}\xi}\sqrt{k}\left(H_{k,n}-\xi\right) + \frac{\sqrt{k}}{H_{k,n}}\left(\ln X_{n-k,n} - \ln U(n/k)\right) \\
&= -\frac{\ln U(n/k)}{H_{k,n}\xi}\sqrt{k}\left(\tilde{H}_{k,n} - E(\tilde{H}_{k,n})\right) \\
&\quad + \frac{1}{H_{k,n}}\left(\frac{\ln U(n/k)}{\xi}+1\right)\sqrt{k}\left(\ln X_{n-k,n} - \ln U(n/k)\right) \\
&\quad - \frac{\ln U(n/k)}{H_{k,n}\xi}\frac{\sqrt{k}B(n/k)}{1+\xi\beta}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\sqrt{k}\left(\ln \hat{A}_{k,n} - \ln A\right) &= -\frac{\ln U(n/k)}{H_{k,n}\xi}\sqrt{k}\left(\tilde{H}_{k,n} - E(\tilde{H}_{k,n})\right) \\
&\quad + \frac{1}{H_{k,n}}\left(\frac{\ln U(n/k)}{\xi}+1\right)\sqrt{k}\left(\ln X_{n-k,n} - \ln U(n/k)\right)
\end{aligned}$$

$$-\frac{1}{\xi}\left(\frac{\ln U(n/k)}{H_{k,n}(1+\beta\xi)}+\frac{1}{\beta}\right)\sqrt{k}B(n/k).$$

Using the fact that $\ln U(n/k)/\ln(n/k)\to\xi$ as $n/k\to\infty$, the result now follows from Lemma 2.1 and Corollary 3.4 in Hsing (1991). □

Proof of Theorem 3.2. Using the approach from Beirlant et al. (2009), we have with $\kappa_n=\kappa(X_{n-k,n})$

$$\begin{aligned}
k^{-1/2}Z_{k,n} &:= \frac{n}{k}\bar{F}(X_{n-k,n})-1\\
&= \frac{A}{(k/n)X_{n-k,n}^{1/\xi}}\left(1+\frac{\kappa_n}{\xi}(1+o_p(1))\right)-1\\
&= \frac{A}{\hat{A}_{k,n}^{EP}}X_{n-k,n}^{1/\hat{\xi}_{k,n}-1/\xi}\left(1+\frac{\kappa_n}{\xi}(1+o_p(1))\right)\left(1-\frac{\hat{\kappa}_{k,n}}{\hat{\xi}_{k,n}}\right)-1\\
&= \frac{A}{\hat{A}_{k,n}^{EP}}\left(1-\frac{1}{\xi\hat{\xi}_{k,n}}(\hat{\xi}_{k,n}-\xi)\ln X_{n-k,n}(1+o_p(1))\right)\\
&\qquad\times\left(1+\left(\frac{\kappa_n}{\xi}-\frac{\hat{\kappa}_{k,n}}{\hat{\xi}_{k,n}}\right)(1+o_p(1))\right)-1,
\end{aligned}$$

from which it follows, using $\hat{\kappa}_{k,n}-\kappa_n=O_p(k^{-1/2})$ and $\hat{\xi}_{k,n}-\xi=O_p(k^{-1/2})$ from Theorem 3.1 in Beirlant et al. (2009),

$$\begin{aligned}
&\frac{A}{\hat{A}_{k,n}^{EP}}-1\\
&\quad= \frac{k^{-1/2}Z_{k,n}+1}{\left(1-\frac{\hat{\xi}_{k,n}-\xi}{\xi\hat{\xi}_{k,n}}\ln X_{n-k,n}(1+o_p(1))\right)\left(1+\left(\frac{\kappa_n}{\xi}-\frac{\hat{\kappa}_{k,n}}{\hat{\xi}_{k,n}}\right)(1+o_p(1))\right)}-1\\
&\quad= (k^{-1/2}Z_{k,n}+1)\left(1+\frac{1}{\xi\hat{\xi}_{k,n}}(\hat{\xi}_{k,n}-\xi)\ln X_{n-k,n}(1+o_p(1))\right)\\
&\qquad\times\left(1-\left(\frac{\kappa_n}{\xi}-\frac{\hat{\kappa}_{k,n}}{\hat{\xi}_{k,n}}\right)(1+o_p(1))\right)-1.
\end{aligned}$$

This implies that $\sqrt{k}(A/\hat{A}_{k,n}^{EP}-1)$ has the same limit distribution as

$$\frac{\ln U(n/k)}{\xi^2}\sqrt{k}(\hat{\xi}_{k,n}-\xi)-\xi^{-1}\sqrt{k}(\hat{\kappa}_{k,n}-\kappa_n)+Z_{k,n}.$$

From Theorem 3.1 in Beirlant et al. (2009) it follows that this stochastic sum is asymptotically unbiased when $\sqrt{k}B(n/k) \to \lambda$, while the asymptotic variance follows from the variance of $\hat{\xi}_{k,n}$ which has the asymptotic dominating coefficient $\ln U(n/k)/\xi^2$ in this asymptotic representation. □

A.3 The dependence between tests on scale and shape

We now derive the covariance matrix of

$$\left(\frac{\xi\sqrt{k}}{\ln U\left(\frac{n}{k}\right)} (\ln \hat{A}_{k,n} - \ln A), \sqrt{k}\left(\frac{H_{k,n}}{\xi} - 1 \right) \right).$$

From

$$\frac{\ln X_{n-k,n}}{H_{k,n}} - \frac{\ln U\left(\frac{n}{k}\right)}{\xi} = \frac{1}{\xi}\left(\ln X_{n-k,n} - \ln U\left(\frac{n}{k}\right) \right) - \frac{\ln X_{n-k,n}}{\xi H_{k,n}} (H_{k,n} - \xi),$$

we have using the notation from the proof of Theorem 3.1 that

$$\frac{\xi T_{k,n}^{(1)}}{\ln U\left(\frac{n}{k}\right)} = \sqrt{k}\left(\frac{\ln X_{n-k,n} - \ln U\left(\frac{n}{k}\right)}{\ln U\left(\frac{n}{k}\right)} \right) - \sqrt{k}\left(\frac{H_{k,n} - \xi}{H_{k,n}} \right) \frac{\ln X_{n-k,n}}{\ln U(n/k)}$$

$$\sim_p \sqrt{k}\left(\frac{\ln X_{n-k,n} - \ln U\left(\frac{n}{k}\right)}{\ln U\left(\frac{n}{k}\right)} \right) - \sqrt{k}\frac{H_{k,n} - \xi}{\xi}.$$

We hence have concerning the asymptotic covariance

$$Acov\left(\frac{\xi T_{k,n}^{(1)}}{\ln U\left(\frac{n}{k}\right)}, \sqrt{k}\frac{H_{k,n} - \xi}{\xi} \right)$$

$$= Acov\left(\sqrt{k}\frac{\ln X_{n-k,n} - \ln U\left(\frac{n}{k}\right)}{\ln U\left(\frac{n}{k}\right)}, \sqrt{k}\frac{H_{k,n} - \xi}{\xi} \right) - Avar\left(\sqrt{k}\frac{H_{k,n} - \xi}{\xi} \right).$$

From (3.11) we know that the asymptotic variance in this expression is asymptotically equal to $1 + \chi + \omega - 2\psi$:

$$Acov\left(\frac{\xi T_{k,n}^{(1)}}{\ln U\left(\frac{n}{k}\right)}, \sqrt{k}\frac{H_{k,n}-\xi}{\xi}\right)$$

$$= \frac{k}{\xi \ln U\left(\frac{n}{k}\right)} Acov\left(\ln X_{n-k,n} - \ln U\left(\frac{n}{k}\right), H_{k,n}-\xi\right) - (1+\chi+\omega-2\psi).$$

Following Hsing (1991), approximating $H_{k,n}$ by $H_{k,n}^{+} - \left(\ln X_{n-k,n} - \ln U\left(\frac{n}{k}\right)\right)$ with $H_{k,n}^{+} = \frac{1}{k}\sum_{j=1}^{k} \max\{\ln X_{n-j+1,n} - \ln U\left(\frac{n}{k}\right), 0\}$, we find

$$k\, Acov\left(\ln X_{n-k,n} - \ln U\left(\frac{n}{k}\right), H_{k,n}-\xi\right)$$

$$\approx k\, Acov\left(\ln X_{n-k,n} - \ln U\left(\frac{n}{k}\right), H_{k,n}^{+}-\xi\right) - k\, Avar\left(\ln X_{n-k,n} - \ln U\left(\frac{n}{k}\right)\right).$$

From Corollary 3.4 in Hsing (1991) it then follows that

$$k\, Acov\left(\ln X_{n-k,n} - \ln U\left(\frac{n}{k}\right), H_{k,n}^{+}-\xi\right) = \xi^2(1+\psi),$$

$$k\, Avar\left(\ln X_{n-k,n} - \ln U\left(\frac{n}{k}\right)\right) = \xi^2(1+\omega),$$

which results in

$$Acov\left(\frac{\xi T_{k,n}^{(1)}}{\ln U\left(\frac{n}{k}\right)}, \sqrt{k}\frac{H_{k,n}-\xi}{\xi}\right) = \frac{\xi}{\ln U\left(\frac{n}{k}\right)}(\psi-\omega) - (1+\chi+\omega-2\psi).$$

Since $T_{k,n}^{(2)}$ is deterministic it does not play a role in the calculation of the covariance matrix. We then get

$$Acov\left(\frac{\xi\sqrt{k}}{\ln U\left(\frac{n}{k}\right)}(\ln \hat{A}_{k,n} - \ln A), \sqrt{k}\frac{H_{k,n}-\xi}{\xi}\right) = -(1+\chi+\omega-2\psi) + \frac{\xi}{\ln U\left(\frac{n}{k}\right)}(\psi-\omega).$$

Using the obtained expression for the asymptotic variance of both components (see Theorem 3.1 and (3.11)) and the fact that $\ln U\left(\frac{n}{k}\right)/\ln(n/k) \to \xi$ as $n/k \to \infty$ gives the asymptotic covariance matrix of $\left(\frac{\xi\sqrt{k}}{\ln U\left(\frac{n}{k}\right)}(\ln \hat{A}_{k,n} - \ln A), \sqrt{k}\left(\frac{H_{k,n}}{\xi} - 1\right)\right)$:

$$\begin{pmatrix} 1+\chi+\omega-2\psi & -(1+\chi+\omega-2\psi) + \frac{\psi-\omega}{\ln\left(\frac{n}{k}\right)} \\ -(1+\chi+\omega-2\psi) + \frac{\psi-\omega}{\ln\left(\frac{n}{k}\right)} & 1+\chi+\omega-2\psi \end{pmatrix}$$

$$= (1+\chi+\omega-\psi)\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \frac{\psi-\omega}{\ln\left(\frac{n}{k}\right)}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \qquad (A.2)$$

A.4 3D plots of P-values for tests

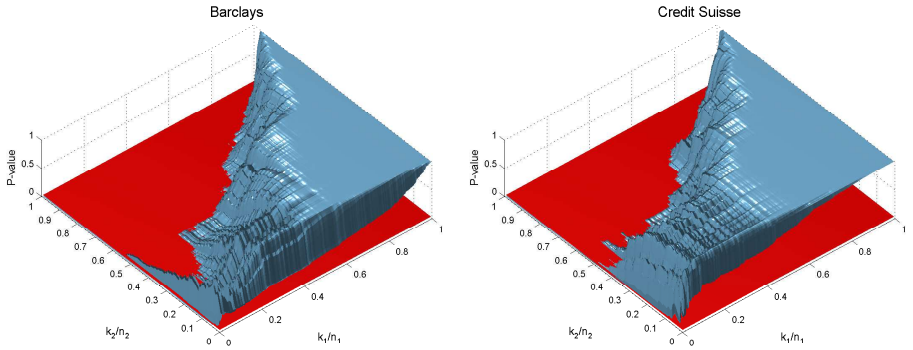


Figure A.1: P-values for testing differences in shape using $T^{(\xi)}_{k_1,k_2,n_1,n_2}$ for all possible choices of k_1 and k_2 for pre- and post-crisis negative log-returns for Barclays and Credit Suisse.

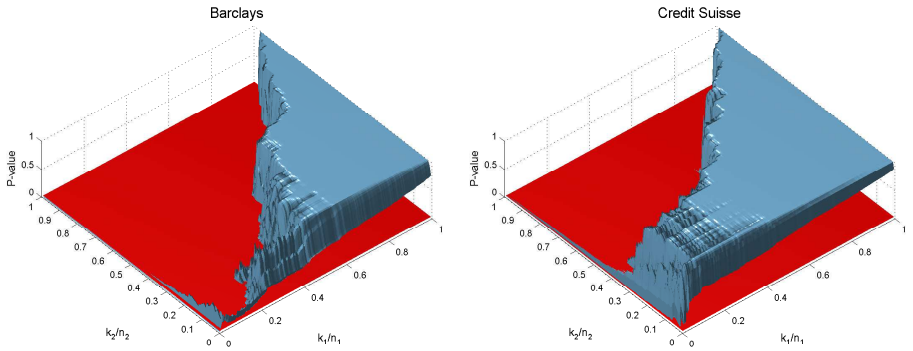


Figure A.2: P-values for testing differences in scale using $T^{(A)}_{k_1,k_2,n_1,n_2}$ for all possible choices of k_1 and k_2 for pre- and post-crisis negative log-returns for Barclays and Credit Suisse.

Appendix B

Appendix for Chapter 4

B.1 Proofs for Section 4.3.3

Proposition B.1. *Under the condition of Theorem 4.1, one can define a sequence of Brownian motions* $\{W_n(s) \mid s > 0\}$*, such that for* $\varepsilon > 0$

(a)
$$\max_{j=1,\dots,k} \left(\frac{j}{k+1}\right)^{0.5+\varepsilon} \left| \sqrt{k} \left[\frac{X_{n-j+1,n} - U_T\left(\frac{n+1}{k+1}\right)}{a_{T,k,n}} - \frac{1}{\xi}\left(\left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}} \right)^{-\xi} - 1 \right) \right] \right.$$
$$+ \frac{b_{T,k,n}}{1+b_{T,k,n}} \left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}} \right)^{-1-\xi} W_n\left(\frac{j}{k+1}\right)$$
$$\left. + \sqrt{k} A\left(\frac{1}{\bar{F}_Y(T)(1+b_{T,k,n})} \right) \Psi_{\xi,\rho}\left(\frac{1+b_{T,k,n}}{1+\frac{j}{k+1}b_{T,k,n}} \right) \right| \to_p 0$$

(b)
$$\max_{j=1,\dots,k} \left(\frac{j}{k+1}\right)^{0.5+\varepsilon} \left| \sqrt{k} \left[\frac{X_{n-j+1,n} - X_{n-k,n}}{a_{T,k,n}} - \frac{1}{\xi}\left(\left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}} \right)^{-\xi} - 1 \right) \right] \right.$$
$$+ \frac{b_{T,k,n}}{1+b_{T,k,n}} \left[\left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}} \right)^{-1-\xi} W_n\left(\frac{j}{k+1}\right) - W_n(1) \right]$$
$$\left. + \sqrt{k} A\left(\frac{1}{\bar{F}_Y(T)(1+b_{T,k,n})} \right) \Psi_{\xi,\rho}\left(\frac{1+b_{T,k,n}}{1+\frac{j}{k+1}b_{T,k,n}} \right) \right| \to_p 0.$$

Proof. In order to derive (a), note that for $j = 1, \ldots, k$,

$$X_{n-j+1,n} - U_T\left(\frac{n+1}{k+1}\right) =_d U_T(Y_{n-j+1,n}) - U_T\left(\frac{n+1}{k+1}\right)$$

$$= U_Y\left(\frac{1+b_{T,k,n}}{1+\frac{1}{Y_{n-j+1,n}D_T}}\frac{1}{\bar{F}_Y(T)(1+b_{T,k,n})}\right) - U_Y\left(\frac{1}{\bar{F}_Y(T)(1+b_{T,k,n})}\right)$$

where we used (4.6), and where $Y_{1,n} \leq Y_{2,n} \leq \ldots \leq Y_{n,n}$ denote the order statistics of an i.i.d. sample from a standard Pareto distribution with distribution function $1 - 1/x$ for $x \geq 1$. Hence, using (4.27) with

$$t = \frac{1}{\bar{F}_Y(T)(1+b_{T,k,n})} \quad \text{and} \quad x = \frac{1+b_{T,k,n}}{1+\frac{n+1}{jY_{n-j+1,n}}\frac{j}{k+1}b_{T,k,n}},$$

we obtain

$$\frac{X_{n-j+1,n} - U_T\left(\frac{n+1}{k+1}\right)}{a_{T,k,n}} = \frac{1}{\xi}\left(\left(\frac{1+\frac{n+1}{jY_{n-j+1,n}}\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}}\right)^{-\xi} - 1\right)$$

$$+ A\left(\frac{1}{\bar{F}_Y(T)(1+b_{T,k,n})}\right)\Psi_{\xi,\rho}\left(\frac{1+b_{T,k,n}}{1+\frac{j}{k+1}b_{T,k,n}}\right) + o_p(1). \tag{B.1}$$

Using Lemma 2.4.10 in de Haan and Ferreira (2006) applied to the standard Pareto distribution one gets

$$\max_{j=1,\ldots,k}\left(\frac{j}{k+1}\right)^{0.5+\varepsilon}\left|\sqrt{k}\left(Y_{n-j+1,n}\frac{j}{n} - 1\right) - \left(\frac{j}{k+1}\right)^{-1} W_n\left(\frac{j}{k+1}\right)\right| \to_p 0.$$

Using the mean value theorem we now obtain

$$\frac{1}{\xi}\left(\left(\frac{1+\frac{n+1}{jY_{n-j+1,n}}\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}}\right)^{-\xi} - \left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}}\right)^{-\xi}\right)$$

$$= \frac{b_{T,k,n}}{1+b_{T,k,n}}\frac{j}{k+1}\left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}}\right)^{-1-\xi}\left(\frac{jY_{n-j+1,n}}{n} - 1\right)(1+o_p(1)).$$

Hence, combining this with (B.1) and the result from Lemma 2.4.10 in de Haan and Ferreira (2006), we arrive at (a). Combining (a) with the analogous result for $j = k+1$, one arrives at (b). To this end note that $\Psi_{\xi,\rho}(1) = 0$. □

Proof of Theorem 4.1. This proof follows the approach of the proof of Theorem 3.4.2 in de Haan and Ferreira (2006). Let $\hat{\tau}_k a_{T,k,n} = \hat{\tau}_k^s$, and

$$Z_{T,k,n}\left(\frac{j}{k+1}\right) = \frac{b_{T,k,n}}{1+b_{T,k,n}}\left(\left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}}\right)^{-1-\xi} W_n\left(\frac{j}{k+1}\right) - W_n(1)\right)$$
$$+\sqrt{k}A\left(\frac{1}{\bar{F}_Y(T)(1+b_{T,k,n})}\right)\Psi_{\xi,\rho}\left(\frac{1+b_{T,k,n}}{1+\frac{j}{k+1}b_{T,k,n}}\right).$$

Then, uniformly in $j \in \{1,\ldots,k\}$,

$$1+\hat{\tau}_k^s\frac{E_{j,k}}{a_{T,k,n}} = \left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}}\right)^{-\xi} + \frac{1}{\xi}(\hat{\tau}_k^s-\xi)\left(\left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}}\right)^{-\xi}-1\right)$$
$$+\hat{\tau}_k^s\frac{1}{\sqrt{k}}Z_{T,k,n}\left(\frac{j}{k+1}\right)+o_p(1).$$

Using $\ln(1+u) = u(1+o(1))$ if $u \downarrow 0$, we get

$$\ln\left(\left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}}\right)^{\xi}\left(1+\hat{\tau}_k^s\frac{E_{j,k}}{a_{T,k,n}}\right)\right) = \frac{1}{\xi}(\hat{\tau}_k^s-\xi)\left(1-\left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}}\right)^{\xi}\right)$$
$$+\hat{\tau}_k^s\frac{1}{\sqrt{k}}Z_{T,k,n}\left(\frac{j}{k+1}\right)\left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}}\right)^{\xi}+o_p(1).$$

Hence, the first term on the left hand side of (4.15) is given by

$$\frac{1}{k-1}\sum_{j=2}^{k}\ln(1+\hat{\tau}_k E_{j,k})$$
$$=\left[-\xi\int_0^1\ln\left(\frac{1+ub_{T,k,n}}{1+b_{T,k,n}}\right)du+\frac{1}{\xi}(\hat{\tau}_k^s-\xi)\int_0^1\left(1-\left(\frac{1+ub_{T,k,n}}{1+b_{T,k,n}}\right)^{\xi}\right)du\right.$$
$$\left.+\hat{\tau}_k^s\frac{1}{\sqrt{k}}\int_0^1 Z_{T,k,n}(u)\left(\frac{1+ub_{T,k,n}}{1+b_{T,k,n}}\right)^{\xi}du\right]$$

$$\sim \left[\xi\left(1-\frac{\ln(1+b_{T,k,n})}{b_{T,k,n}}\right)+\frac{1}{\xi}(\hat{\tau}_k^s-\xi)\left(1-\frac{1+b_{T,k,n}}{b_{T,k,n}(1+\xi)}\left(1-(1+b_{T,k,n})^{-1-\xi}\right)\right)\right.$$
$$\left.+\hat{\tau}_k^s\frac{1}{\sqrt{k}}\int_0^1 Z_{T,k,n}(u)\left(\frac{1+ub_{T,k,n}}{1+b_{T,k,n}}\right)^{\xi}du\right]. \tag{B.2}$$

Moreover, using Proposition B.1(b) with $j=1$, we obtain

$$\left(1+\hat{\tau}_k^s\frac{E_{1,k}}{a_{T,k,n}}\right)^{-1/\hat{\xi}_k}=\left(1+\hat{\tau}_k^s\frac{1}{\xi}\left((1+b_{T,k,n})^{\xi}-1\right)+\hat{\tau}_k^s\frac{1}{\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)\right)^{-1/\hat{\xi}_k}$$
$$=(1+b_{T,k,n})^{-\frac{\xi}{\hat{\xi}_k}}\left(1+(\hat{\tau}_k^s-\xi)\frac{1}{\xi}\left(1-(1+b_{T,k,n})^{-\xi}\right)\right.$$
$$\left.+\hat{\tau}_k^s\frac{1}{\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)(1+b_{T,k,n})^{-\xi}\right)^{-1/\hat{\xi}_k}$$
$$=(1+b_{T,k,n})^{-1}\left(1+(\hat{\xi}_k-\xi)\frac{1}{\xi}\ln(1+b_{T,k,n})-(\hat{\tau}_k^s-\xi)\frac{1}{\xi^2}\left(1-(1+b_{T,k,n})^{-\xi}\right)\right.$$
$$\left.-\frac{\hat{\tau}_k^s}{\hat{\xi}_k}\frac{1}{\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)(1+b_{T,k,n})^{-\xi}\right)(1+o_p(1))$$

where we used the series expansions

$$e^{-\left(\frac{\xi}{\hat{\xi}_k}-1\right)\ln(1+b_{T,n,k})}=1-\left(\frac{\xi}{\hat{\xi}_k}-1\right)\ln(1+b_{T,n,k})(1+o_p(1))$$

and $(1+u)^{-1/\xi}=1-\frac{1}{\xi}u(1+o(1))$. Hence, the second term on the left hand side of (4.15) equals

$$-\hat{\xi}_k\frac{\left(1+\hat{\tau}_k^s\frac{E_{1,k}}{a_{T,k,n}}\right)^{-1/\hat{\xi}_k}\ln\left(1+\hat{\tau}_k^s\frac{E_{1,k}}{a_{T,k,n}}\right)^{-1/\hat{\xi}_k}}{1-\left(1+\hat{\tau}_k^s\frac{E_{1,k}}{a_{T,k,n}}\right)^{-1/\hat{\xi}_k}}$$
$$=-\hat{\xi}_k(1+b_{T,k,n})^{-1}\left(1+\frac{(\hat{\xi}_k-\xi)}{\xi}\ln(1+b_{T,k,n})-\frac{(\hat{\tau}_k^s-\xi)}{\xi^2}\left(1-(1+b_{T,k,n})^{-\xi}\right)\right.$$
$$\left.-\frac{\hat{\tau}_k^s}{\hat{\xi}_k}\frac{1}{\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)(1+b_{T,k,n})^{-\xi}\right)$$

$$\times \ln(1+b_{T,k,n})^{-1} \times \left(1 - \frac{(\hat{\xi}_k - \xi)}{\xi} + \frac{(\hat{\tau}_k^s - \xi)}{\xi^2} \frac{\left(1-(1+b_{T,k,n})^{-\xi}\right)}{\ln(1+b_{T,k,n})} \right.$$

$$\left. + \frac{\hat{\tau}_k^s}{\hat{\xi}_k} \frac{1}{\sqrt{k}} Z_{T,k,n}\left(\frac{1}{k+1}\right) \frac{(1+b_{T,k,n})^{-\xi}}{\ln(1+b_{T,k,n})} \right)$$

$$/ \left[\frac{b_{T,k,n}}{1+b_{T,k,n}} \left(1 - \frac{(\hat{\xi}_k - \xi)}{\xi} \frac{\ln(1+b_{T,k,n})}{b_{T,k,n}} + \frac{(\hat{\tau}_k^s - \xi)}{\xi^2} \frac{\left(1-(1+b_{T,k,n})^{-\xi}\right)}{b_{T,k,n}} \right.\right.$$

$$\left.\left. + \frac{\hat{\tau}_k^s}{\hat{\xi}_k} \frac{1}{\sqrt{k}} Z_{T,k,n}\left(\frac{1}{k+1}\right) \frac{(1+b_{T,k,n})^{-\xi}}{b_{T,k,n}} \right) \right] (1+o_p(1))$$

$$\sim \left[\hat{\xi}_k \frac{\ln(1+b_{T,k,n})}{b_{T,k,n}} + (\hat{\xi}_k - \xi) \frac{\ln(1+b_{T,k,n})}{b_{T,k,n}} \left(-1 + \frac{1+b_{T,k,n}}{b_{T,k,n}} \ln(1+b_{T,k,n})\right) \right.$$

$$- \frac{(\hat{\tau}_k^s - \xi)}{\xi} \left(1-(1+b_{T,k,n})^{-\xi}\right) \left(\frac{1+b_{T,k,n}}{b_{T,k,n}} - \frac{1}{\ln(1+b_{T,k,n})}\right) \frac{\ln(1+b_{T,k,n})}{b_{T,k,n}}$$

$$\left. - \hat{\tau}_k^s \frac{1}{\sqrt{k}} Z_{T,k,n}\left(\frac{1}{k+1}\right) (1+b_{T,k,n})^{-\xi} \left(\frac{1+b_{T,k,n}}{b_{T,k,n}} - \frac{1}{\ln(1+b_{T,k,n})}\right) \frac{\ln(1+b_{T,k,n})}{b_{T,k,n}} \right]. \tag{B.3}$$

Combining (4.15), (B.2) and (B.3) gives

$$\left[\xi \left(1 - \frac{\ln(1+b_{T,k,n})}{b_{T,k,n}}\right) + \frac{1}{\xi}(\hat{\tau}_k^s - \xi) \left(1 - \frac{1+b_{T,k,n}}{b_{T,k,n}(1+\xi)} \left(1-(1+b_{T,k,n})^{-1-\xi}\right)\right) \right.$$

$$\left. + \hat{\tau}_k^s \frac{1}{\sqrt{k}} \int_0^1 Z_{T,k,n}(u) \left(\frac{1+u b_{T,k,n}}{1+b_{T,k,n}}\right)^{\xi} du \right] (1+o_p(1))$$

$$+ \left[\hat{\xi}_k \frac{\ln(1+b_{T,k,n})}{b_{T,k,n}} + (\hat{\xi}_k - \xi) \frac{\ln(1+b_{T,k,n})}{b_{T,k,n}} \left(-1 + \frac{1+b_{T,k,n}}{b_{T,k,n}} \ln(1+b_{T,k,n})\right) \right.$$

$$- \frac{(\hat{\tau}_k^s - \xi)}{\xi} \left(1-(1+b_{T,k,n})^{-\xi}\right) \left(\frac{1+b_{T,k,n}}{b_{T,k,n}} - \frac{1}{\ln(1+b_{T,k,n})}\right) \frac{\ln(1+b_{T,k,n})}{b_{T,k,n}}$$

$$\left. - \hat{\tau}_k^s \frac{1}{\sqrt{k}} Z_{T,k,n}\left(\frac{1}{k+1}\right) (1+b_{T,k,n})^{-\xi} \left(\frac{1+b_{T,k,n}}{b_{T,k,n}} - \frac{1}{\ln(1+b_{T,k,n})}\right) \frac{\ln(1+b_{T,k,n})}{b_{T,k,n}} \right]$$

$$= \hat{\xi}_k (1+o_p(1)).$$

This equation can be written as

$$\left[(\hat{\xi}_k-\xi)\left(-1+\frac{(1+b_{T,k,n})\ln^2(1+b_{T,k,n})}{b_{T,k,n}^2}\right)\right.$$

$$+\frac{1}{\xi}(\hat{\tau}_k^s-\xi)\left(\frac{\xi}{1+\xi}\frac{1+b_{T,k,n}}{b_{T,k,n}}\left(1-(1+b_{T,k,n})^{-1-\xi}\right)\right.$$

$$\left.-\frac{(1+b_{T,k,n})}{b_{T,k,n}^2}\ln(1+b_{T,k,n})\left(1-(1+b_{T,k,n})^{-\xi}\right)\right)$$

$$+\frac{\hat{\tau}_k^s}{\sqrt{k}}\int_0^1 Z_{T,k,n}(u)\left(\frac{1+ub_{T,k,n}}{1+b_{T,k,n}}\right)^{\xi}du$$

$$\left.-\frac{\hat{\tau}_k^s}{\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)(1+b_{T,k,n})^{-\xi}\left(\frac{(1+b_{T,k,n})\ln(1+b_{T,k,n})}{b_{T,k,n}^2}-\frac{1}{b_{T,k,n}}\right)\right](1+o_p(1))$$

$$=0. \tag{B.4}$$

The left hand side of (4.16) yields, using similar asymptotic methods as above,

$$\frac{1}{k-1}\sum_{j=2}^{k}\left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}}\right)^{\xi}\left[1-\frac{(\hat{\tau}_k^s-\xi)}{\xi}\left(1-\left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}}\right)^{\xi}\right)\right.$$

$$\left.-\frac{\hat{\tau}_k^s}{\sqrt{k}}Z_{T,k,n}\left(\frac{j}{k+1}\right)\left(1-\left(\frac{1+\frac{j}{k+1}b_{T,k,n}}{1+b_{T,k,n}}\right)^{\xi}\right)\right](1+o_p(1))$$

$$=\left[\frac{1}{1+\xi}\frac{1+b_{T,k,n}}{b_{T,k,n}}\left(1-(1+b_{T,k,n})^{-1-\xi}\right)\right.$$

$$-\frac{(\hat{\tau}_k^s-\xi)}{\xi}\left(\frac{\xi(1+b_{T,k,n})}{(1+\xi)(1+2\xi)}-\frac{(1+b_{T,k,n})^{-\xi}}{1+\xi}+\frac{(1+b_{T,k,n})^{-2\xi}}{1+2\xi}\right)$$

$$\left.-\frac{\hat{\tau}_k^s}{\sqrt{k}}\int_0^1 Z_{T,k,n}(u)\left(\frac{1+ub_{T,k,n}}{1+b_{T,k,n}}\right)^{2\xi}du\right](1+o_p(1)). \tag{B.5}$$

The right hand side of (4.16) is asymptotically equivalent to (where we used again Proposition B.1(b) with $j=1$)

$$
\begin{aligned}
&\frac{1}{1+\hat{\xi}_k}\left[1-\frac{1}{1+b_{T,k,n}}\left(1+(\hat{\xi}_k-\xi)\frac{1}{\xi}\ln(1+b_{T,k,n})-\frac{1}{\xi^2}(\hat{\tau}_k^s-\xi)\left(1-(1+b_{T,k,n})^{-\xi}\right)\right.\right.\\
&\qquad\left.-\hat{\tau}_k^s\frac{1}{\hat{\xi}_k\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)(1+b_{T,k,n})^{-\xi}\right)\\
&\quad\times\left.\left((1+b_{T,k,n})^{\xi}+\frac{1}{\xi}(\hat{\tau}_k^s-\xi)\left((1+b_{T,k,n})^{\xi}-1\right)+\hat{\tau}_k^s\frac{1}{\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)\right)^{-1}\right]\\
&\times\left[1-\frac{1}{1+b_{T,k,n}}\left(1+(\hat{\xi}_k-\xi)\frac{1}{\xi}\ln(1+b_{T,k,n})-\frac{1}{\xi^2}(\hat{\tau}_k^s-\xi)\left(1-(1+b_{T,k,n})^{-\xi}\right)\right.\right.\\
&\quad\left.\left.-\hat{\tau}_k^s\frac{1}{\hat{\xi}_k\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)(1+b_{T,k,n})^{-\xi}\right)\right]^{-1}\\
&\sim\frac{1}{1+\hat{\xi}_k}\frac{(1+b_{T,k,n})\left(1-(1+b_{T,k,n})^{-1-\xi}\right)}{b_{T,k,n}}\\
&\times\left(1-(\hat{\xi}_k-\xi)\frac{1}{\xi}\ln(1+b_{T,k,n})\frac{(1+b_{T,k,n})^{-1-\xi}}{1-(1+b_{T,k,n})^{-1-\xi}}\right.\\
&\qquad+(\hat{\tau}_k^s-\xi)\frac{1+\xi}{\xi^2}\frac{(1+b_{T,k,n})^{-1-\xi}}{1-(1+b_{T,k,n})^{-1-\xi}}\left(1-(1+b_{T,k,n})^{-\xi}\right)\\
&\qquad\left.+\frac{\hat{\tau}_k^s}{\hat{\xi}_k}\frac{1}{\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)\frac{1+\hat{\xi}_k}{\hat{\xi}_k}\frac{(1+b_{T,k,n})^{-1-2\xi}}{1-(1+b_{T,k,n})^{-1-\xi}}\right)\\
&\times\left[1-(\hat{\xi}_k-\xi)\frac{1}{\xi}\frac{\ln(1+b_{T,k,n})}{b_{T,k,n}}+(\hat{\tau}_k^s-\xi)\frac{1}{\xi^2}\frac{1-(1+b_{T,k,n})^{-\xi}}{b_{T,k,n}}\right.\\
&\qquad\left.+\frac{\hat{\tau}_k^s}{\hat{\xi}_k}\frac{1}{\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)\frac{(1+b_{T,k,n})^{-\xi}}{b_{T,k,n}}\right]^{-1}\\
&\sim\frac{1}{1+\hat{\xi}_k}\frac{(1+b_{T,k,n})\left(1-(1+b_{T,k,n})^{-1-\xi}\right)}{b_{T,k,n}}\\
&\quad+\frac{(\hat{\xi}_k-\xi)}{\xi(1+\xi)}\frac{1+b_{T,k,n}}{b_{T,k,n}^2}\ln(1+b_{T,k,n})\left(1-(1+b_{T,k,n})^{-\xi}\right)
\end{aligned}
$$

$$+ (\hat{\tau}_k^s - \xi)\frac{1}{\xi^2(1+\xi)}\frac{1+b_{T,k,n}}{b_{T,k,n}}\left(1-(1+b_{T,k,n})^{-\xi}\right)\left(-\frac{1-(1+b_{T,k,n})^{-1-\xi}}{b_{T,k,n}}\right.$$

$$\left.+ (1+\xi)(1+b_{T,k,n})^{-1-\xi}\right)$$

$$- \frac{1}{1+\xi}\frac{\hat{\tau}_k^s}{\hat{\xi}_k}\frac{1}{\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)\frac{(1+b_{T,k,n})^{1-\xi}}{b_{T,k,n}^2}\left(1-(1+b_{T,k,n})^{-\xi}\right). \tag{B.6}$$

Combining (4.16), (B.5) and (B.6) leads to (after some lengthy calculations)

$$(\hat{\xi}_k - \xi)\frac{1}{\xi}\frac{1+b_{T,k,n}}{b_{T,k,n}}\left(\frac{\xi}{1+\xi}\left(1-(1+b_{T,k,n})^{-1-\xi}\right) - \frac{\ln(1+b_{T,k,n})}{b_{T,k,n}}\left(1-(1+b_{T,k,n})^{-\xi}\right)\right)$$

$$- (\hat{\tau}_k^s - \xi)\frac{1+b_{T,k,n}}{b_{T,k,n}}\frac{1}{\xi}\left(\frac{\xi}{1+2\xi}\left(1-(1+b_{T,k,n})^{-1-2\xi}\right) - \frac{1}{b_{T,k,n}}\frac{1}{\xi}\left(1-(1+b_{T,k,n})^{-\xi}\right)^2\right)$$

$$= \frac{\xi(\xi+1)}{\sqrt{k}}\int_0^1 Z_{T,k,n}(u)\left(\frac{1+ub_{T,k,n}}{1+b_{T,k,n}}\right)^{2\xi}du$$

$$- \frac{1}{\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)\frac{(1+b_{T,k,n})^{1-\xi}}{b_{T,k,n}^2}\left(1-(1+b_{T,k,n})^{-\xi}\right). \tag{B.7}$$

□

Proof of Theorem 4.2.

$$\hat{Q}_{T,k}(1-p)$$

$$= X_{n-k,n} + \frac{1}{\hat{\tau}_k}\left(\left(1+\frac{k}{n\hat{D}_T}\right)^{\hat{\xi}_k}\left(1+\frac{1}{d_n}\frac{k}{n\hat{D}_T}\right)^{-\hat{\xi}_k} - 1\right)$$

$$= X_{n-k,n} + \frac{1}{\hat{\tau}_k}\left(\left(\frac{1-\frac{1}{k}}{(1+\hat{\tau}_k E_{1,k})^{-\frac{1}{\hat{\xi}_k}} - \frac{1}{k}}\right)^{\hat{\xi}_k}\left(1+\frac{1}{d_n}\frac{k}{n\hat{D}_T}\right)^{-\hat{\xi}_k} - 1\right)$$

$$= X_{n-k,n} + \frac{1}{\hat{\tau}_k}\left((1+\hat{\tau}_k E_{1,k})\left(\frac{1-\frac{1}{k}}{1-\frac{1}{k}(1+\hat{\tau}_k E_{1,k})^{\frac{1}{\hat{\xi}_k}}}\right)^{\hat{\xi}_k}\left(1+\frac{1}{d_n}\frac{k}{n\hat{D}_T}\right)^{-\hat{\xi}_k} - 1\right)$$

$$= X_{n-k,n} + \frac{1}{\hat{\tau}_k}\left((1+\hat{\tau}_k E_{1,k})\left(1 - \frac{\hat{\xi}_k}{k}\left(1-(1+\hat{\tau}_k E_{1,k})^{\frac{1}{\hat{\xi}_k}}\right)(1+o_p(1))\right)\right.$$
$$\left.\times\left(1 - \frac{\hat{\xi}_k}{d_n}\frac{k}{n\hat{D}_T}(1+o_p(1))\right) - 1\right)$$
$$= X_{n-k,n} + \left(E_{1,k} + (1+\hat{\tau}_k E_{1,k})\left(-\frac{\hat{\xi}_k}{\hat{\tau}_k k}\left(1-(1+\hat{\tau}_k E_{1,k})^{\frac{1}{\hat{\xi}_k}}\right)\right.\right.$$
$$\left.\left.-\frac{\hat{\xi}_k}{d_n\hat{\tau}_k}\frac{k}{n\hat{D}_T}\right)(1+o_p(1))\right)$$
$$= X_{n,n} - \frac{\hat{\xi}_k}{\hat{\tau}_k}(1+\hat{\tau}_k E_{1,k})\left(\frac{1}{k}\left(1-(1+\hat{\tau}_k E_{1,k})^{\frac{1}{\hat{\xi}_k}}\right) + \frac{1}{d_n}\frac{k}{n\hat{D}_T}\right)(1+o_p(1)).$$

Hence,

$$\hat{Q}_{T,k}(1-p) - Q_T(1-p)$$
$$= \left(X_{n,n} - Q_T\left(1-\frac{1}{n}\right)\right) + \left(Q_T\left(1-\frac{1}{n}\right) - Q_T(1-p)\right)$$
$$- \frac{\hat{\xi}_k}{\hat{\tau}_k}(1+\hat{\tau}_k E_{1,k})\left(\frac{1}{k}\left(1-(1+\hat{\tau}_k E_{1,k})^{\frac{1}{\hat{\xi}_k}}\right) + \frac{1}{d_n}\frac{k}{n\hat{D}_T}\right)\left(1+o_p\left(\frac{1}{d_n}\right)\right). \tag{B.8}$$

First, using again the notation $Y_{1,n} \leq Y_{2,n} \leq \ldots \leq Y_{n,n}$ for the order statistics of an i.i.d. sample of size n from a standard Pareto distribution, we obtain using (4.28)

$$X_{n,n} - Q_T\left(1-\frac{1}{n}\right) =_d U_T(Y_{n,n}) - U_T(n)$$
$$= U_Y\left(\frac{1}{\bar{F}_Y(T)\left(1+\frac{n}{Y_{n,n}}\frac{1}{nD_T}\right)}\right) - U_Y\left(\frac{1}{\bar{F}_Y(T)\left(1+\frac{1}{nD_T}\right)}\right)$$
$$= U_Y\left(\frac{1+\frac{1}{nD_T}}{1+\frac{n}{Y_{n,n}}\frac{1}{nD_T}}\frac{1}{\bar{F}_Y(T)\left(1+\frac{1}{nD_T}\right)}\right) - U_Y\left(\frac{1}{\bar{F}_Y(T)\left(1+\frac{1}{nD_T}\right)}\right)$$

$$
\begin{aligned}
&= a_Y\left(\frac{1}{\bar{F}_Y(T)\left(1+\frac{1}{k}b_{T,k,n}\right)}\right)\left(\frac{1}{\xi}\left(\left(\frac{1+\frac{b_{T,k,n}}{k}}{1+\frac{n}{Y_{n,n}}\frac{b_{T,k,n}}{k}}\right)^{\xi}-1\right)\right.\\
&\quad\left.+A\left(\frac{1}{\bar{F}_Y(T)\left(1+\frac{1}{k}b_{T,k,n}\right)}\right)\Psi_{\xi,\rho}\left(\frac{1+\frac{b_{T,k,n}}{k}}{1+\frac{n}{Y_{n,n}}\frac{b_{T,k,n}}{k}}\right)\right)(1+o_p(1))\\
&= a_Y\left(\frac{1}{\bar{F}_Y(T)}\right)\left(1+\frac{b_{T,k,n}}{k}\right)^{-\xi}\left(1+A\left(\frac{1}{\bar{F}_Y(T)}\right)C\left(\frac{\left(1+\frac{b_{T,k,n}}{k}\right)^{-\rho}-1}{\rho}\right)\right)\\
&\quad\times\left(-b_{T,k,n}\frac{1}{k}\left(\frac{n}{Y_{n,n}}-1\right)\left(1+O_p\left(\frac{1}{k}\right)\right)+O_p\left(\frac{1}{k^2}\right)\right)\\
&= a_Y\left(\frac{1}{\bar{F}_Y(T)}\right)\left(1-\frac{\xi b_{T,k,n}}{k}-A\left(\frac{1}{\bar{F}_Y(T)}\right)C\frac{b_{T,k,n}}{k}+O_p\left(\frac{1}{k^2}\right)\right)\\
&\quad\times\left(-\frac{b_{T,k,n}}{k}(E-1)+O_p\left(\frac{1}{k^2}\right)\right). \qquad \text{(B.9)}
\end{aligned}
$$

Here, we used that $\frac{n}{Y_{n,n}} =_d E+O_p\left(\frac{1}{n}\right)$ and that $\Psi_{\xi,\rho}\left(1+\frac{D}{k}\right)=O\left(\frac{1}{k^2}\right)$ for any constant D. Furthermore,

$$
\begin{aligned}
&Q_T\left(1-\frac{1}{n}\right)-Q_T\left(1-p\right)\\
&\quad= U_Y\left(\frac{1}{\bar{F}_Y(T)\left(1+\frac{1}{nD_T}\right)}\right)-U_Y\left(\frac{1}{\bar{F}_Y(T)\left(1+\frac{p}{D_T}\right)}\right)\\
&\quad= U_Y\left(\frac{1+\frac{b_{T,k,n}}{d_n}}{1+\frac{b_{T,k,n}}{k}}\frac{1}{\bar{F}_Y(T)\left(1+\frac{b_{T,k,n}}{d_n}\right)}\right)-U_Y\left(\frac{1}{\bar{F}_Y(T)\left(1+\frac{p}{D_T}\right)}\right)\\
&\quad= a_Y\left(\frac{1}{\bar{F}_Y(T)\left(1+\frac{b_{T,k,n}}{d_n}\right)}\right)\left(\frac{1}{\xi}\left(\left(\frac{1+\frac{b_{T,k,n}}{d_n}}{1+\frac{b_{T,k,n}}{k}}\right)^{\xi}-1\right)\right.\\
&\qquad\left.+A\left(\frac{1}{\bar{F}_Y(T)\left(1+\frac{b_{T,k,n}}{d_n}\right)}\right)\Psi_{\xi,\rho}\left(\frac{1+\frac{b_{T,k,n}}{d_n}}{1+\frac{b_{T,k,n}}{k}}\right)\right)(1+o_p(1))
\end{aligned}
$$

$$= a_Y\left(\frac{1}{\bar{F}_Y(T)}\right)\left(1+\frac{b_{T,k,n}}{d_n}\right)^{-\xi}\left(1+A\left(\frac{1}{\bar{F}_Y(T)}\right)C\left(\frac{\left(1+\frac{b_{T,k,n}}{d_n}\right)^{-\rho}-1}{\rho}\right)\right)$$

$$\times\left(b_{T,k,n}\left(\frac{1}{d_n}-\frac{1}{k}\right)\left(1+O\left(\frac{1}{d_n}\right)\right)+O\left(\frac{1}{d_n^2}\right)\right)$$

$$= a_Y\left(\frac{1}{\bar{F}_Y(T)}\right)\left(b_{T,k,n}\left(\frac{1}{d_n}-\frac{1}{k}\right)+O\left(\frac{1}{d_n^2}\vee\frac{1}{k^2}\right)\right). \tag{B.10}$$

Finally, using $k/(n\hat{D}_T) = \left((1+\hat{\tau}_k E_{1,k})^{\frac{1}{\hat{\xi}_k}}-1\right)(1+O_p(1/k))$ and derivations as in the proof of Theorem 4.1, the third term in the right hand side of (B.8) equals

$$-\left(\frac{\hat{\xi}_k}{\hat{\tau}_k}\frac{1}{a_{T,k,n}}\right)a_{T,k,n}(1+\hat{\tau}_k E_{1,k})\left(\frac{1}{k}\left(1-(1+\hat{\tau}_k E_{1,k})^{\frac{1}{\hat{\xi}_k}}\right)+\frac{1}{d_n}\frac{k}{n\hat{D}_T}\right)$$

$$= -a_Y\left(\frac{1}{\bar{F}_Y(T)}\right)(1+b_{T,k,n})^{-\xi}\left(1+A\left(\frac{1}{\bar{F}_Y(T)}\right)C\left(\frac{(1+b_{T,k,n})^{-\rho}-1}{\rho}\right)\right)$$

$$\times\left(1+\left(\frac{\hat{\xi}_k}{\hat{\tau}_k}\frac{1}{a_{T,k,n}}-1\right)\right)$$

$$\times(1+b_{T,k,n})^{\xi}\left(1+(\hat{\tau}_k^s-\xi)\frac{1}{\xi}\left(1-(1+b_{T,k,n})^{-\xi}\right)\right.$$

$$\left.+\frac{\hat{\tau}_k^s}{\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)(1+b_{T,k,n})^{-\xi}\right)$$

$$\times\left(\left(1-(1+\hat{\tau}_k E_{1,k})^{\frac{1}{\hat{\xi}_k}}\right)\left(\frac{1}{d_n}-\frac{1}{k}\right)+O_p\left(\frac{1}{d_n k}\right)\right)$$

$$= -a_Y\left(\frac{1}{\bar{F}_Y(T)}\right)\left(1+A\left(\frac{1}{\bar{F}_Y(T)}\right)C\left(\frac{(1+b_{T,k,n})^{-\rho}-1}{\rho}\right)\right)\left(1+O_p\left(\frac{1}{k}\right)\right)$$

$$\times\left(1+\left(\frac{\hat{\xi}_k}{\hat{\tau}_k}\frac{1}{a_{T,k,n}}-1\right)\right)$$

$$\times\left(1+(\hat{\tau}_k^s-\xi)\frac{1}{\xi}\left(1-(1+b_{T,k,n})^{-\xi}\right)+\frac{\hat{\tau}_k^s}{\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)(1+b_{T,k,n})^{-\xi}\right)$$

$$\times b_{T,k,n}\left(\frac{1}{d_n}-\frac{1}{k}\right)$$

$$\times\Bigg(1-(\hat{\xi}_k-\xi)\frac{1}{\xi}\frac{1+b_{T,k,n}}{b_{T,k,n}}\ln(1+b_{T,k,n})$$

$$+(\hat{\tau}_k^s-\xi)\frac{1}{\xi^2}\frac{1+b_{T,k,n}}{b_{T,k,n}}\left(1-(1+b_{T,k,n})^{-\xi}\right)$$

$$+\frac{1}{\sqrt{k}}Z_{T,k,n}\left(\frac{1}{k+1}\right)\frac{(1+b_{T,k,n})^{1-\xi}}{b_{T,k,n}}\Bigg). \tag{B.11}$$

The result follows from joining (B.8), (B.9), (B.10) and (B.11) and retaining terms of order $O\left(\frac{1}{k}\right)$, $O\left(\left(\frac{1}{d_n}-\frac{1}{k}\right)A\left(\frac{1}{\bar{F}_Y(T)}\right)\right)$ and $O\left(\left(\frac{1}{d_n}-\frac{1}{k}\right)\frac{1}{\sqrt{k}}\right)$. □

Proof of Theorem 4.3. Note that using (4.5) and $\bar{F}_Y(T)=D_T F_Y(T)$, we obtain

$$T_{k,n}=k\left(1+\hat{\tau}_k^s\frac{E_{1,k}}{a_{T,k,n}}\right)^{-1/\hat{\xi}_k}$$

$$=k\left(1+\hat{\tau}_k^s\frac{U_T(Y_{n,n})-U_T(Y_{n-k,n})}{a_{T,k,n}}\right)^{-1/\hat{\xi}_k}$$

$$=k\Bigg(1+\frac{\hat{\tau}_k^s}{a_Y\left(\frac{1}{D_T(1+b_{T,k,n})F_Y(T)}\right)}$$

$$\times\left[U_Y\left(\frac{Y_{n,n}}{F_Y(T)(1+Y_{n,n}D_T)}\right)-U_Y\left(\frac{Y_{n-k,n}}{F_Y(T)(1+Y_{n-k,n}D_T)}\right)\right]\Bigg)^{-1/\hat{\xi}_k}$$

$$=k\Bigg(1+\frac{\hat{\tau}_k^s}{a_Y\left(\frac{n/k}{(1+nD_T/k)F_Y(T)}\right)}$$

$$\times\Bigg[U_Y\left(\frac{\frac{Y_{n,n}}{Y_{n-k,n}}}{\frac{1+Y_{n,n}D_T}{1+Y_{n-k,n}D_T}}\frac{\frac{kY_{n-k,n}}{n}\frac{n}{k}}{F_Y(T)\left(1+\frac{kY_{n-k,n}}{n}\frac{nD_T}{k}\right)}\right)$$

$$-U_Y\left(\frac{\frac{kY_{n-k,n}}{n}\frac{n}{k}}{F_Y(T)\left(1+\frac{kY_{n-k,n}}{n}\frac{nD_T}{k}\right)}\right)\Bigg]\Bigg)^{-1/\hat{\xi}_k}.$$

Now one applies (4.27) with $t = \frac{\frac{kY_{n-k,n}}{n}\frac{n}{k}}{F_Y(T)\left(1+\frac{kY_{n-k,n}}{n}\frac{nD_T}{k}\right)} = \frac{n}{k}(1+o_p(1))$ and $x = \frac{Y_{n,n}}{Y_{n-k,n}}\frac{1+Y_{n-k,n}D_T}{1+Y_{n,n}D_T} = U_{1,k}^{-1}(1+o_p(1))$ since $\frac{kY_{n-k,n}}{n} = 1+O_p(1/\sqrt{k})$, $Y_{n,n}/n = 1+o_p(1)$, $nD_T \to 0$ and $Y_{n-k,n}/Y_{n,n} =_d U_{1,k}$, the minimum of an i.i.d. sample of size k from the uniform (0,1) distribution. This, with $\hat{\tau}_k^s/\xi = 1+o_p(1)$, yields

$T_{k,n}$

$$= k\left(1+\frac{\hat{\tau}_k^s}{\xi}\left[U_{1,k}^{-\xi}(1+o_p(1))-1+\xi A\left(\frac{n}{k}(1+o_p(1))\right)\Psi_{\xi,\rho}\left(U_{1,k}^{-1}(1+o_p(1))\right)\right]\right)^{-1/\hat{\xi}_k}$$

$$= k\left(U_{1,k}^{-\xi}(1+o_p(1))+\xi A\left(\frac{n}{k}(1+o_p(1))\right)\Psi_{\xi,\rho}\left(U_{1,k}^{-1}(1+o_p(1))\right)\right)^{-(1/\xi)(1+O_p(1/\sqrt{k}))}$$

$$= kU_{1,k}\left(1+o_p(1)+\xi U_{1,k}^{\xi}A\left(\frac{n}{k}(1+o_p(1))\right)\Psi_{\xi,\rho}(k(1+o_p(1)))\right)^{-1/\xi}$$

because $U_{1,k}^{-1} = O_p(k)$ and $U_{1,k}^{\xi}\Psi_{\xi,\rho}(k(1+o_p(1))) = O_p(1)$. The result now follows from $kU_{1,k} =_d E(1+o_p(1))$. □

B.2 Simulation results

On the next pages, the simulation results as discussed in Section 4.3.2 are shown. The results for the test for truncation and the estimation of ξ are displayed on pages 154 to 157. On pages 158 to 165, the simulation results for the estimation of extreme quantiles are shown.

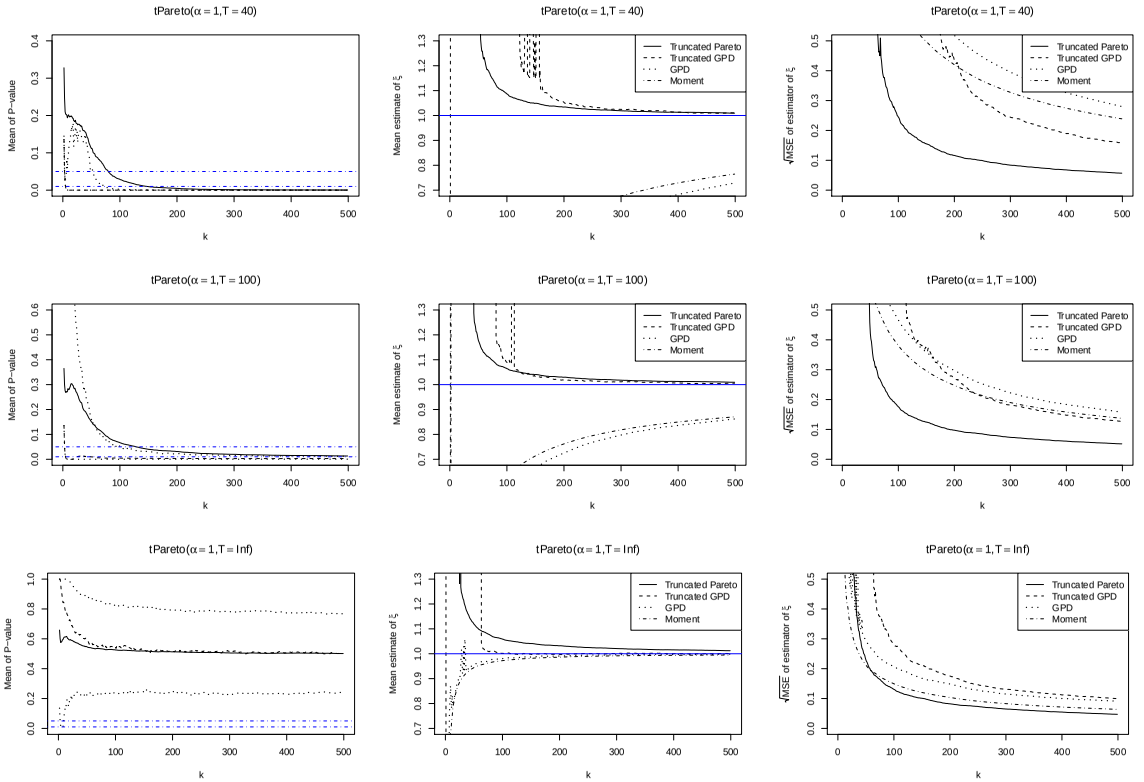


Figure B.1: Means and boxplots of P-values for test (left), means (middle) and root MSE (right) of $\hat{\xi}_k^+$, $\hat{\xi}_k$, $\hat{\xi}_k^\infty$ and $\hat{\xi}_k^{\text{Mom}}$ from the standard Pareto distribution truncated at $Q_Y(0.975)$ (top), $Q_Y(0.99)$ (middle) and not truncated (bottom).

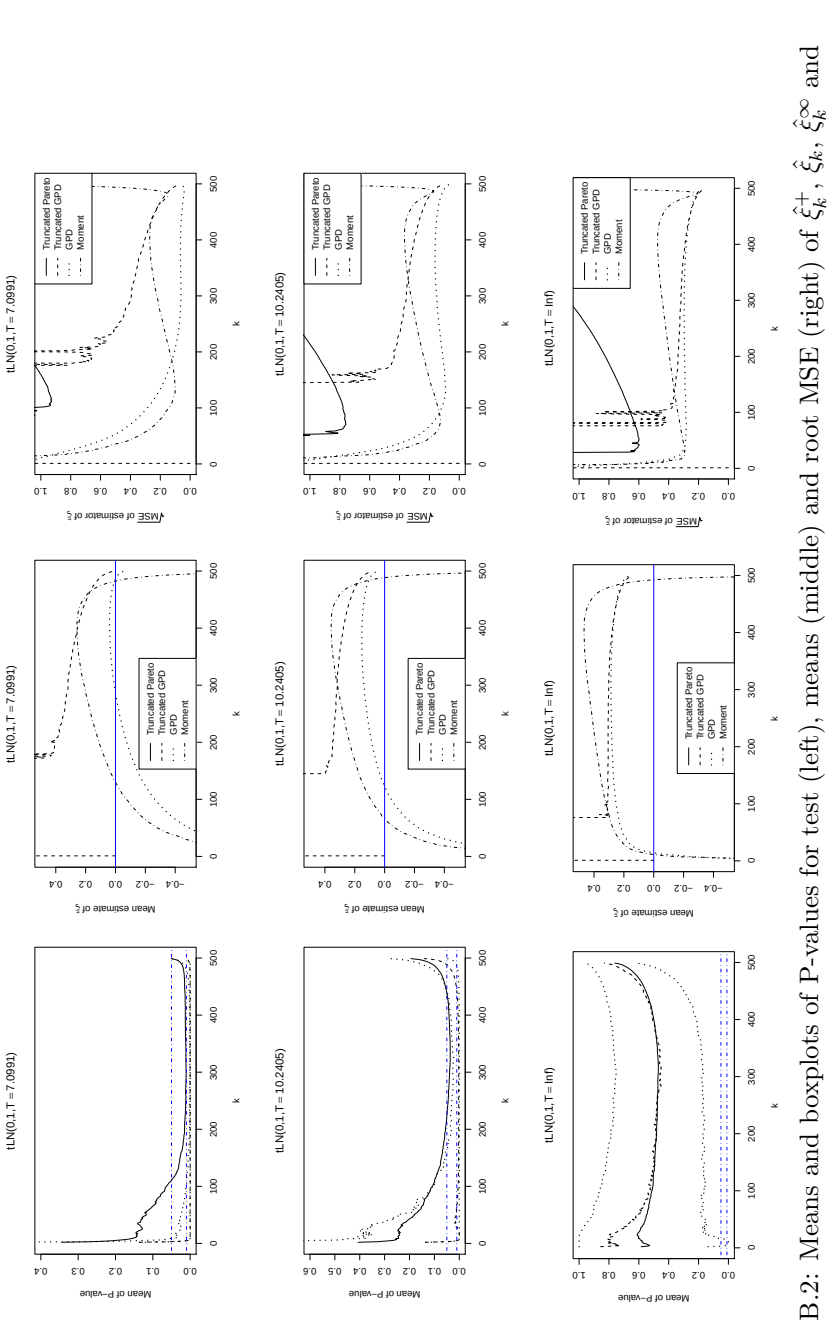


Figure B.2: Means and boxplots of P-values for test (left), means (middle) and root MSE (right) of $\hat{\xi}_k^+$, $\hat{\xi}_k$, $\hat{\xi}_k^\infty$ and $\hat{\xi}_k^{\text{Mom}}$ from the standard lognormal distribution truncated at $Q_Y(0.975)$ (top), $Q_Y(0.99)$ (middle) and not truncated (bottom).

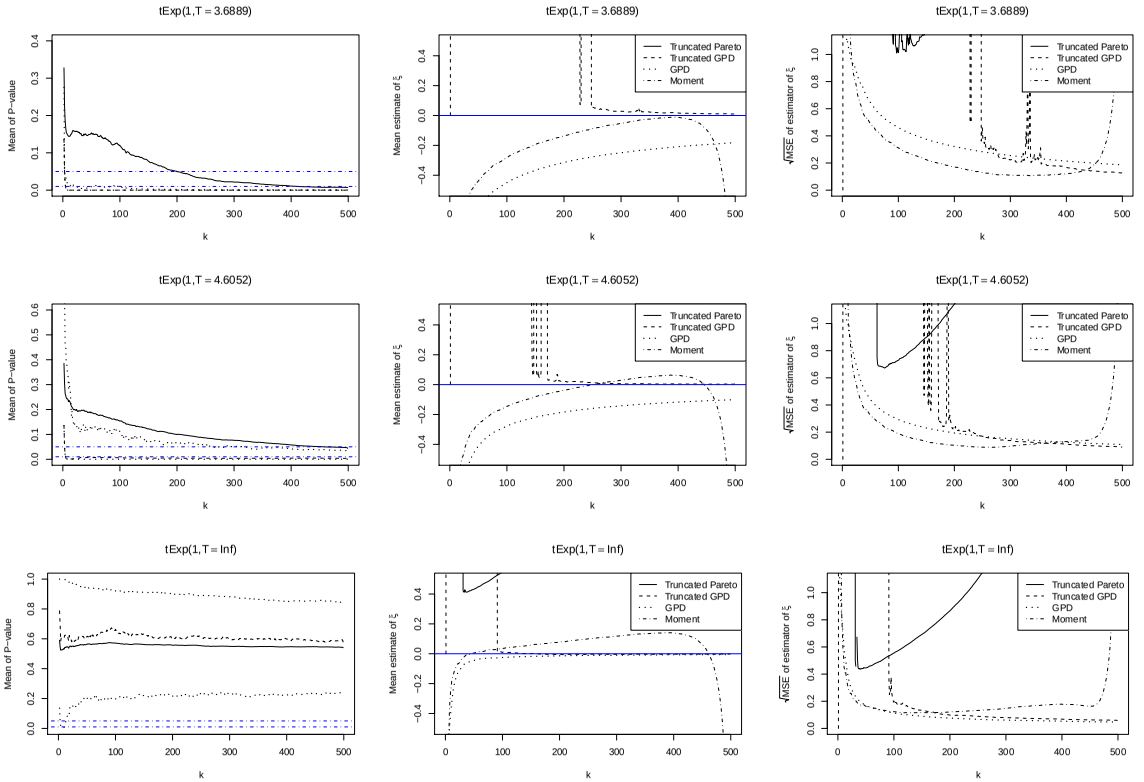


Figure B.3: Means and boxplots of P-values for test (left), means (middle) and root MSE (right) of $\hat{\xi}_k^+$, $\hat{\xi}_k$, $\hat{\xi}_k^\infty$ and $\hat{\xi}_k^{\text{Mom}}$ from the standard exponential distribution truncated at $Q_Y(0.975)$ (top), $Q_Y(0.99)$ (middle) and not truncated (bottom).

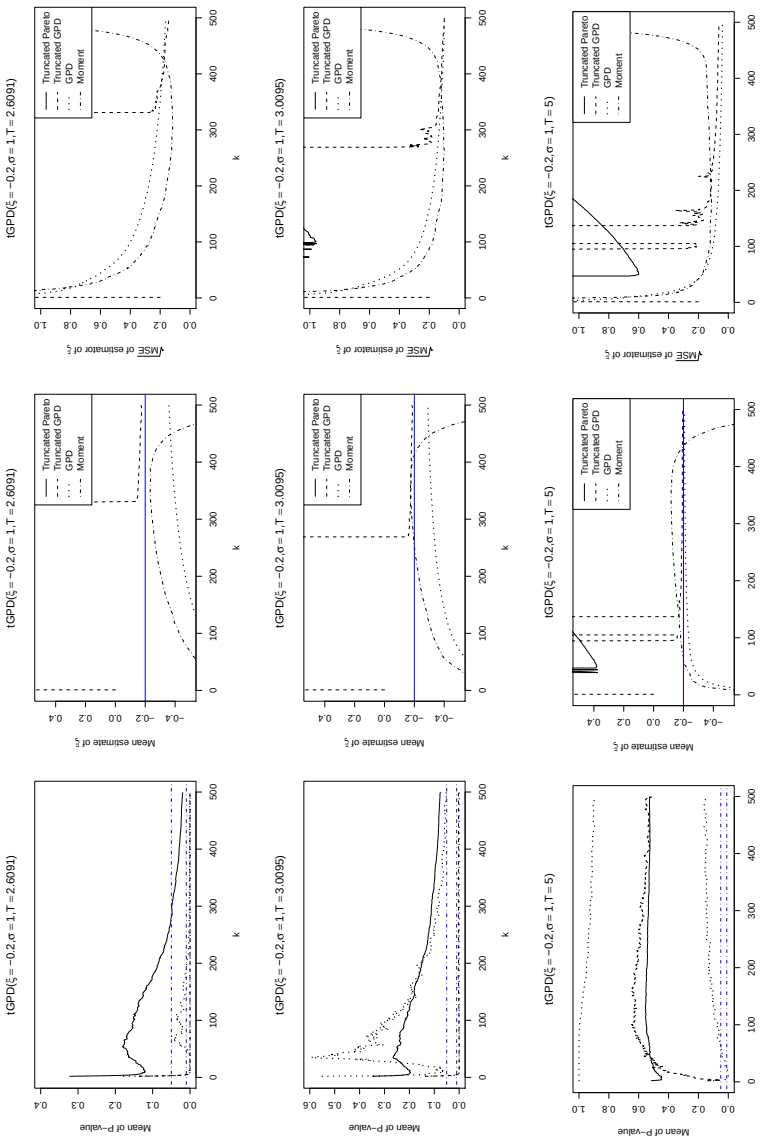


Figure B.4: Means and boxplots of P-values for test (left), means (middle) and root MSE (right) of $\hat{\xi}_k^+$, $\hat{\xi}_k$, $\hat{\xi}_k^\infty$ and $\hat{\xi}_k^{\text{Mom}}$ from GPD(-0.2,1) truncated at $Q_Y(0.975)$ (top), $Q_Y(0.99)$ (middle) and $Q_Y(1)$ (bottom).

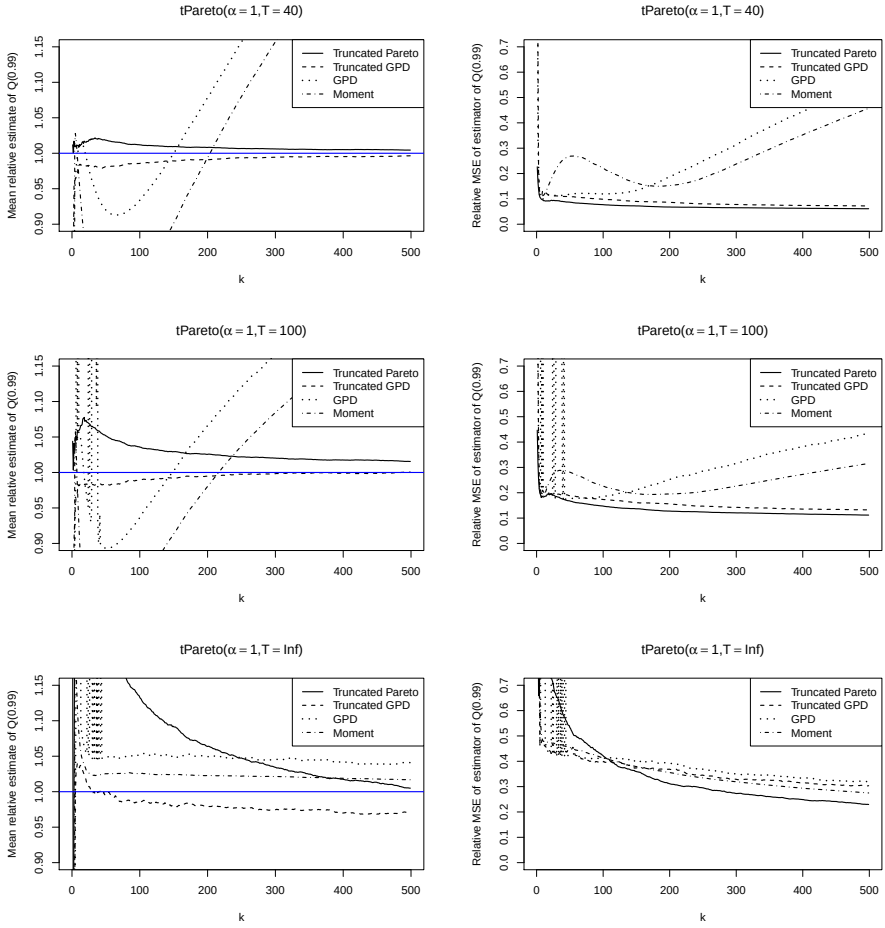


Figure B.5: Mean deviations of $\hat{Q}^{+}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\infty}_{k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\mathrm{Mom}}_{k}(1-p)/Q_T(1-p)$ and corresponding MSE with $p = 0.01$ for the standard Pareto distribution truncated at $Q_Y(0.975)$ (top), $Q_Y(0.99)$ (middle) and not truncated (bottom).

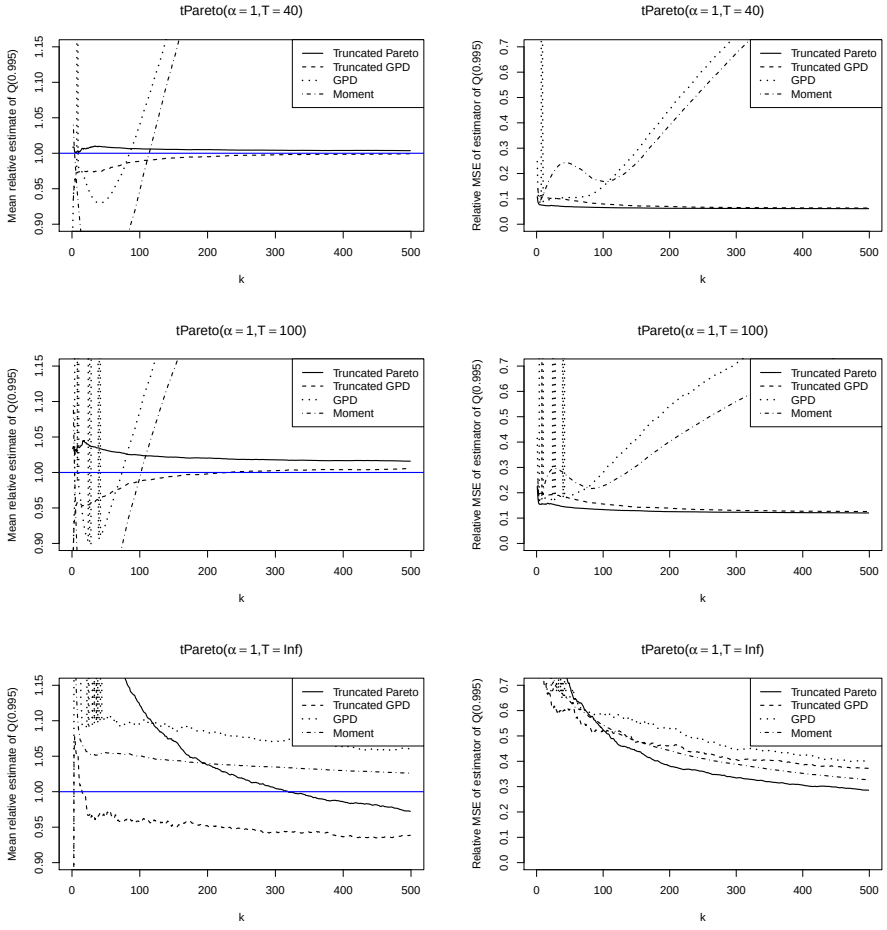


Figure B.6: Mean deviations of $\hat{Q}^{+}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\infty}_{k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\mathrm{Mom}}_{k}(1-p)/Q_T(1-p)$ and corresponding MSE with $p = 0.005$ for the standard Pareto distribution truncated at $Q_Y(0.975)$ (top), $Q_Y(0.99)$ (middle) and not truncated (bottom).

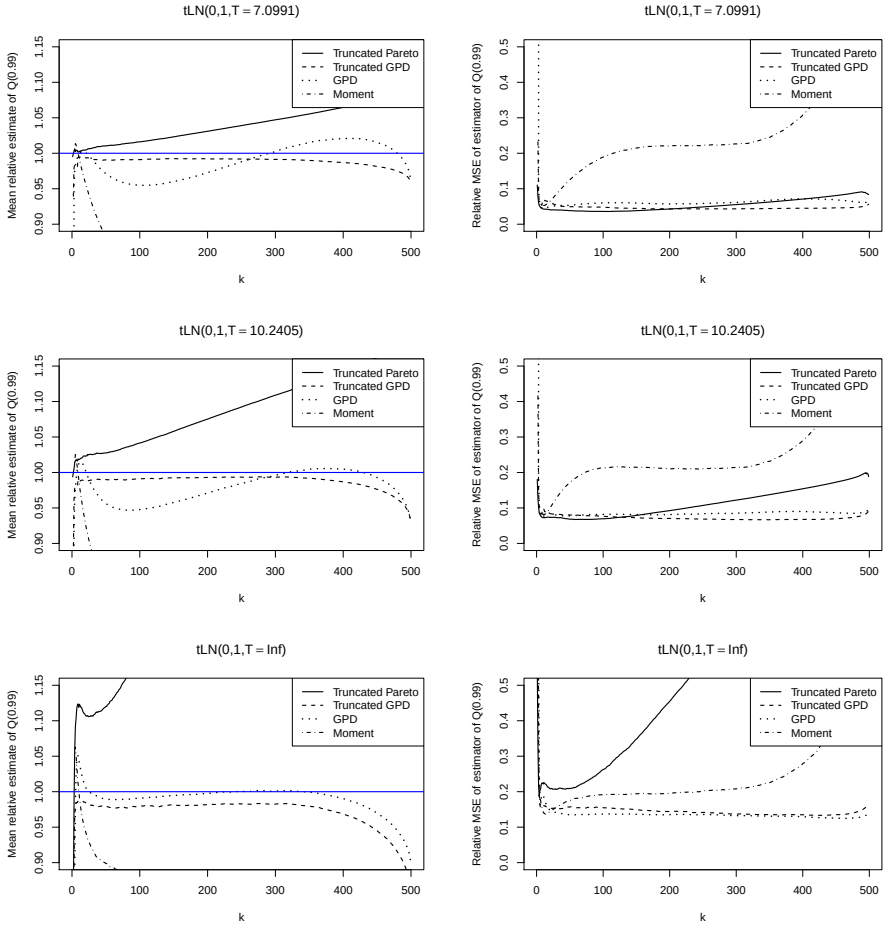


Figure B.7: Mean deviations of $\hat{Q}^{+}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\infty}_{k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\mathrm{Mom}}_{k}(1-p)/Q_T(1-p)$ and corresponding MSE with $p = 0.01$ for the standard lognormal distribution truncated at $Q_Y(0.975)$ (top), $Q_Y(0.99)$ (middle) and not truncated (bottom).

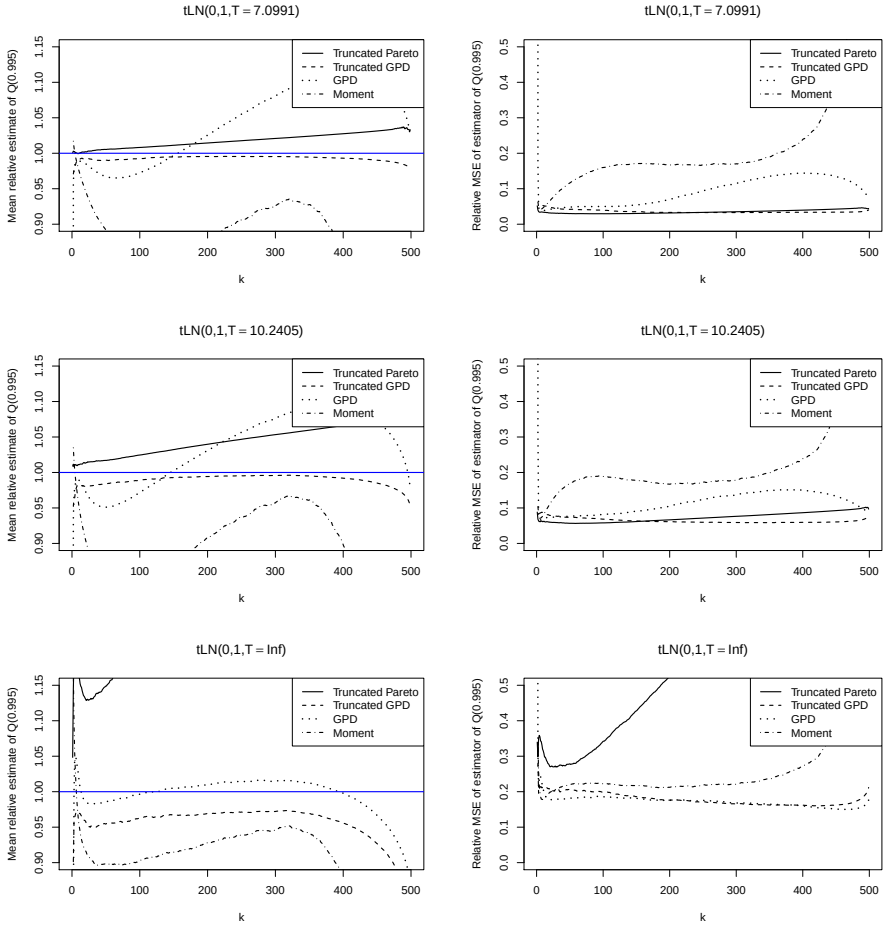


Figure B.8: Mean deviations of $\hat{Q}^{+}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\infty}_{k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\mathrm{Mom}}_{k}(1-p)/Q_T(1-p)$ and corresponding MSE with $p = 0.005$ for the standard lognormal distribution truncated at $Q_Y(0.975)$ (top), $Q_Y(0.99)$ (middle) and not truncated (bottom).

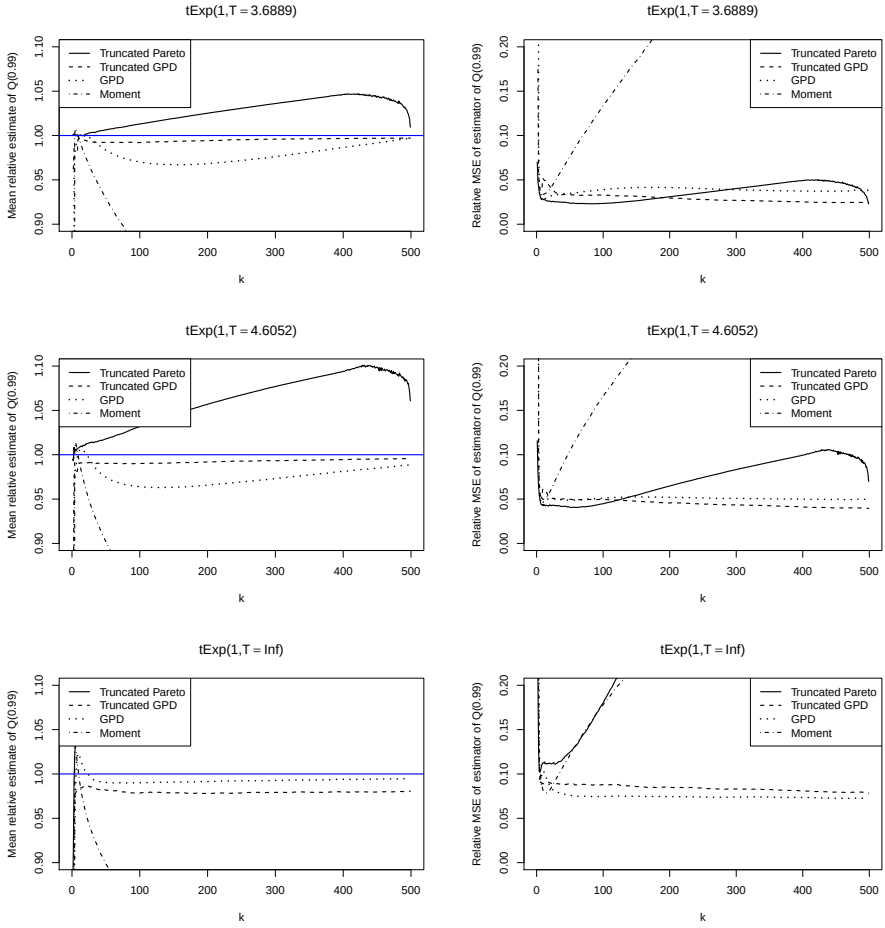


Figure B.9: Mean deviations of $\hat{Q}^{+}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\infty}_{k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\mathrm{Mom}}_{k}(1-p)/Q_T(1-p)$ and corresponding MSE with $p = 0.01$ for the standard exponential distribution truncated at $Q_Y(0.975)$ (top), $Q_Y(0.99)$ (middle) and not truncated (bottom).

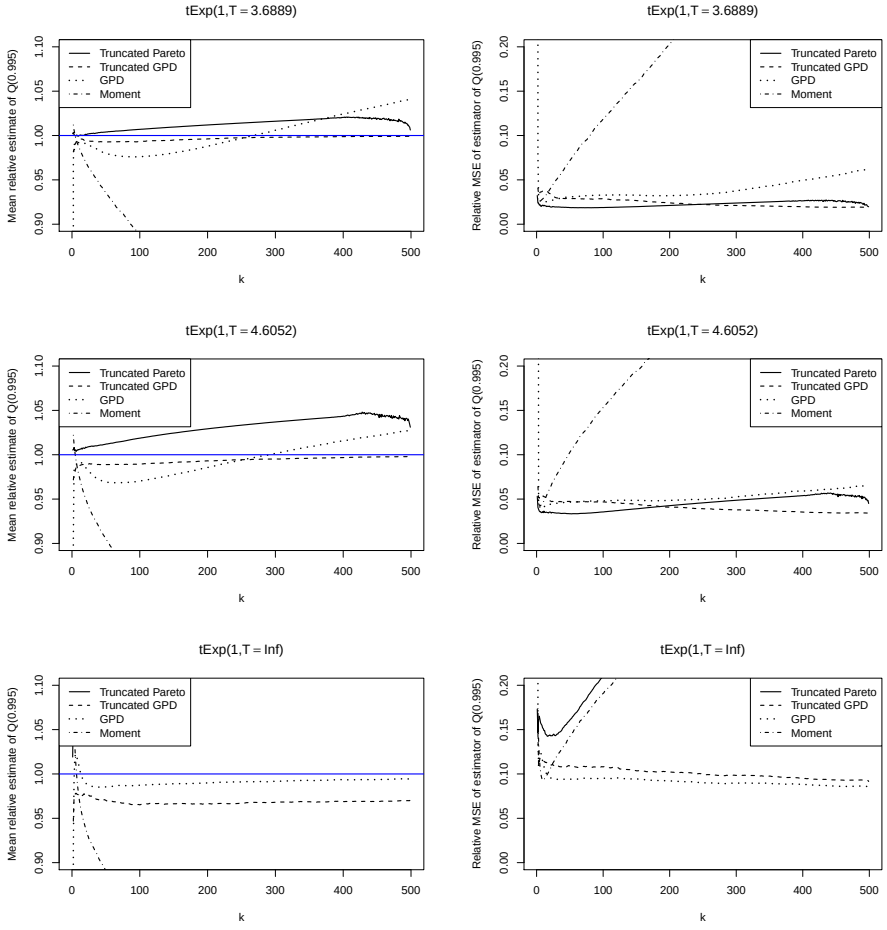


Figure B.10: Mean deviations of $\hat{Q}^{+}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\infty}_{k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\mathrm{Mom}}_{k}(1-p)/Q_T(1-p)$ and corresponding MSE with $p = 0.005$ for the standard exponential distribution truncated at $Q_Y(0.975)$ (top), $Q_Y(0.99)$ (middle) and not truncated (bottom).

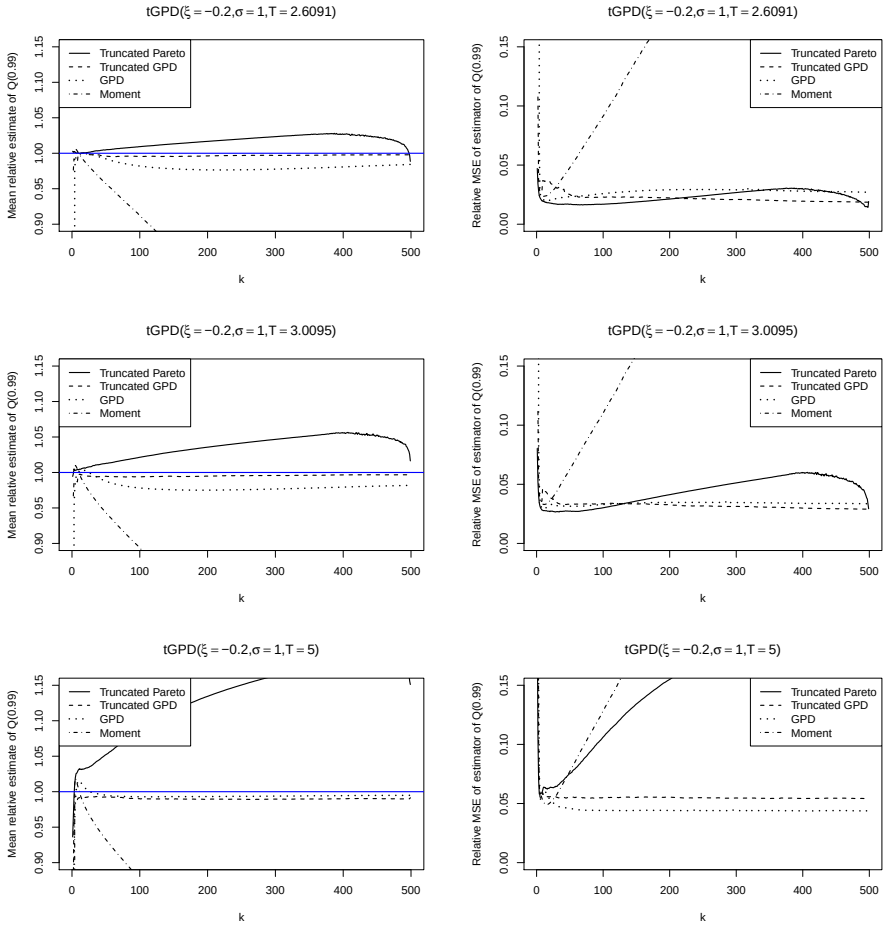


Figure B.11: Mean deviations of $\hat{Q}^{+}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\infty}_{k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\mathrm{Mom}}_{k}(1-p)/Q_T(1-p)$ and corresponding MSE with $p = 0.01$ for GPD(-0.2,1) truncated at $Q_Y(0.975)$ (top), $Q_Y(0.99)$ (middle) and $Q_Y(1)$ (bottom).

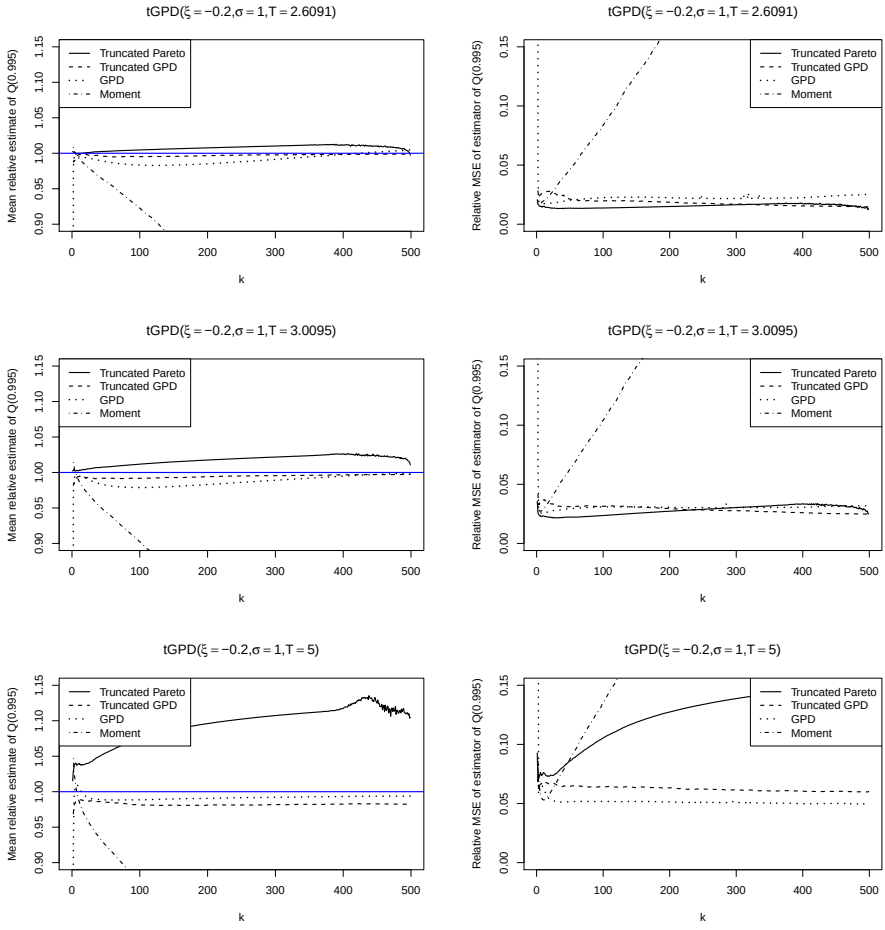


Figure B.12: Mean deviations of $\hat{Q}^{+}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}_{T,k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\infty}_{k}(1-p)/Q_T(1-p)$, $\hat{Q}^{\mathrm{Mom}}_{k}(1-p)/Q_T(1-p)$ and corresponding MSE with $p = 0.005$ for GPD(-0.2,1) truncated at $Q_Y(0.975)$ (top), $Q_Y(0.99)$ (middle) and $Q_Y(1)$ (bottom).

Appendix C

Appendix for Chapter 5

This appendix contains the results for the simulations that are discussed in Section 5.4. For each simulation setting, the relative means and MSE of the endpoint estimates, and coverage percentages of the 90% upper confidence bounds for the endpoint are shown.

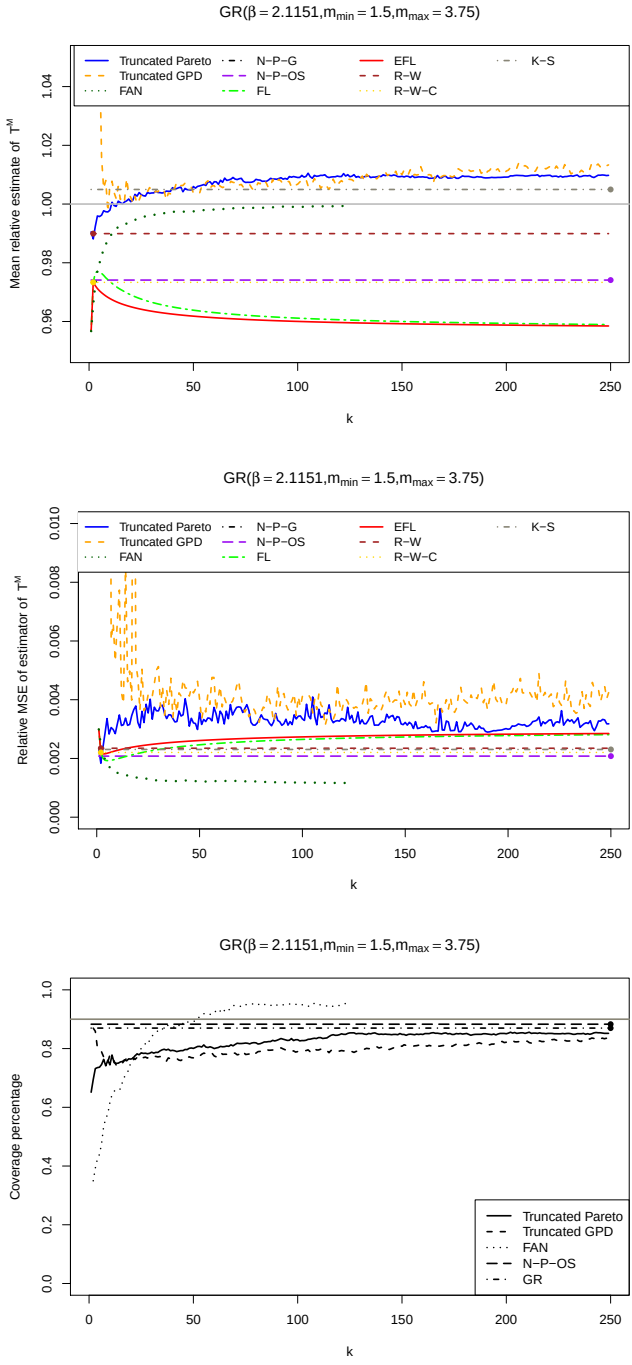


Figure C.1: $GR(\beta = 2.1151, t_M = 1.5, T_M = 3.75)$: relative means of endpoint estimates (top), relative MSE of endpoint estimates (middle) and coverage percentage of 90% upper confidence bounds for the endpoint (bottom).

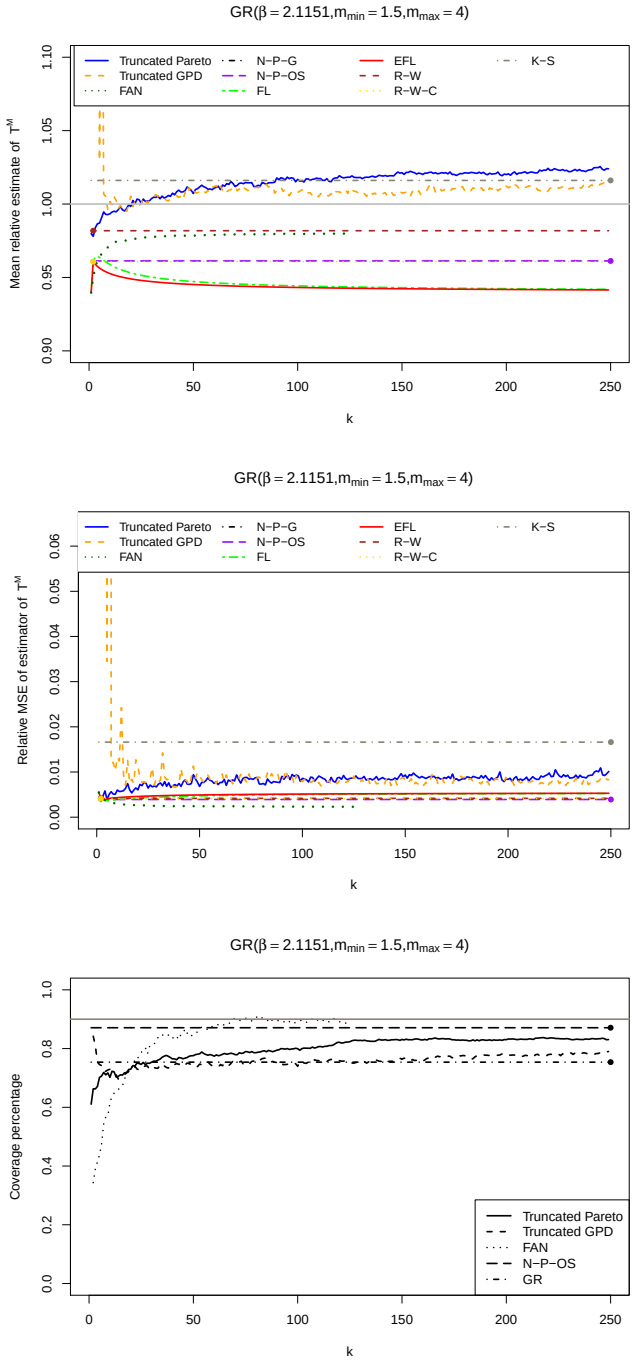


Figure C.2: $GR(\beta = 2.1151, t_M = 1.5, T_M = 4)$: relative means of endpoint estimates (top), relative MSE of endpoint estimates (middle) and coverage percentage of 90% upper confidence bounds for the endpoint (bottom).

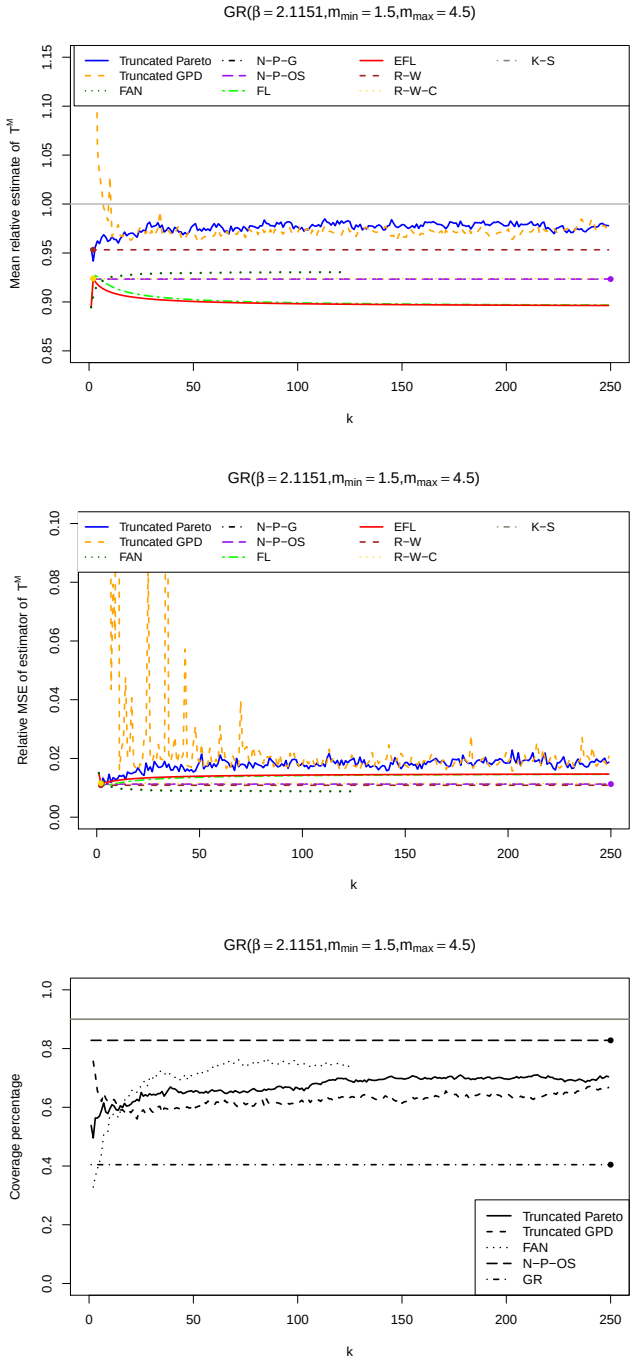


Figure C.3: $GR(\beta = 2.1151, t_M = 1.5, T_M = 4.5)$: relative means of endpoint estimates (top), relative MSE of endpoint estimates (middle) and coverage percentage of 90% upper confidence bounds for the endpoint (bottom).

Appendix D

Appendix for Chapter 6

D.1 Fitting the ME-Pareto model to censored data using the EM algorithm

This appendix contains all the details of the framework described in Section 6.3 applied to the splicing of the ME and Pareto distributions.

D.1.1 Initial step

We estimate the splicing constant π initially by the Turnbull estimator (Turnbull, 1976) in t, i.e. $\pi^{(0)} = \hat{F}_{TB}(t)$ since the estimate for π in the hth iteration ($\pi^{(h)}$) can be seen as the proportion of data points smaller than or equal to t, see (6.10).

Initial values for the ME parameters follow from Verbelen et al. (2016) which are improvements of the starting values from Verbelen et al. (2015). Consider the data $\boldsymbol{d}$ consisting of x_i for all uncensored data points, of l_i for all right censored data points, of u_i for all left censored data points and of the interval midpoints, $\frac{l_i+u_i}{2}$, for all interval censored data points. Restricting the data $\boldsymbol{d}$ to all data points that are smaller than or equal to t gives $\widetilde{\boldsymbol{d}}$ which has length $n_{\widetilde{\boldsymbol{d}}}$. Verbelen et al. (2016) use the following starting values: $\theta^{(0)} = \frac{\max(\widetilde{\boldsymbol{d}})}{s}$, for a

given spread factor s,

$$\boldsymbol{r} = \left(\left\lceil \frac{\hat{Q}(0;\widetilde{\boldsymbol{d}})}{\theta^{(0)}} \right\rceil, \left\lceil \frac{\hat{Q}\left(\frac{1}{M-1};\widetilde{\boldsymbol{d}}\right)}{\theta^{(0)}} \right\rceil, \ldots, \left\lceil \frac{\hat{Q}(1;\widetilde{\boldsymbol{d}})}{\theta^{(0)}} \right\rceil \right) \tag{D.1}$$

and

$$\alpha_j^{(0)} = \frac{\sum_{i=1}^{n_{\widetilde{\boldsymbol{d}}}} I\left(\left\{ r_{j-1}\theta^{(0)} < \widetilde{d}_i \leq r_j\theta^{(0)} \right\}\right)}{n_{\widetilde{\boldsymbol{d}}}}$$

for $j = 1, \ldots, M$, where $r_0 = 0$ for notational convenience and $\hat{Q}(\cdot;\widetilde{\boldsymbol{d}})$ is the empirical quantile function of the data $\widetilde{\boldsymbol{d}}$.

As initial value for ξ, we use the Hill estimator (Hill, 1975) with threshold t applied to $\boldsymbol{d}$.

This gives $\boldsymbol{\Theta}^{(0)} = (\pi^{(0)}, \boldsymbol{\beta}^{(0)}, \theta^{(0)}, \xi^{(0)})$ where $\boldsymbol{\alpha}^{(0)}$ is transformed to $\boldsymbol{\beta}^{(0)}$ using (6.3).

D.1.2 E-step

In the hth iteration of the E-step, we take the conditional expectation of the complete log-likelihood given the incomplete data $\mathcal{X}$ and using the current estimate $\boldsymbol{\Theta}^{(h-1)}$ for $\boldsymbol{\Theta}$. As said before, we assume that there are no shared parameters in π, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$. We can hence split the maximisation in three parts. Therefore, we also consider three parts in the conditional expectation of the complete log-likelihood.

Splicing weight π

We easily obtain the probability (6.9) using (6.4) and (6.5) with $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_1^{(h-1)}$ and $\xi = \xi^{(h-1)}$. This gives

$$P\left(X_i \leq t \,\middle|\, t^l \leq l_i < t < u_i \leq T; \boldsymbol{\Theta}^{(h-1)}\right)$$

$$= \frac{\pi^{(h-1)} - \pi^{(h-1)} F_1\left(l_i; t^l, t, \boldsymbol{r}, \boldsymbol{\Theta}_1^{(h-1)}\right)}{\pi^{(h-1)} + (1-\pi^{(h-1)}) \frac{1-\left(\frac{u_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}}}{1-\left(\frac{T}{t}\right)^{-\frac{1}{\xi^{(h-1)}}}} - \pi^{(h-1)} F_1\left(l_i; t^l, t, \boldsymbol{r}, \boldsymbol{\Theta}_1^{(h-1)}\right)}, \tag{D.2}$$

where

$$F_1\left(l_i; t^l, t, \boldsymbol{r}, \boldsymbol{\Theta}_1^{(h-1)}\right) = \sum_{j=1}^{M} \beta_j^{(h-1)} F_E^t(l_i; t^l, t, r_j, \theta^{(h-1)})$$

$$= \sum_{j=1}^{M} \beta_j^{(h-1)} \frac{F_E(l_i; r_j, \theta^{(h-1)})}{F_E(t; r_j, \theta^{(h-1)}) - F_E(t^l; r_j, \theta^{(h-1)})}.$$

ME distribution

After using arguments similar to those used to obtain (6.8), we get that the conditional expectation of (6.12), the part of the complete data log-likelihood depending on $\boldsymbol{\Theta}_1$, is given by

$$\sum_{i \in S_{i\cdot}} E\left(\sum_{j=1}^{M} Z_{ij} \ln\left(\beta_j f_E^t(X_i; t^l, t, r_j, \theta)\right) \,\middle|\, t^l \le l_i = u_i \le t < T; \boldsymbol{\Theta}_1^{(h-1)}\right)$$

$$+ \sum_{i \in S_{iii\cdot}} E\left(\sum_{j=1}^{M} Z_{ij} \ln\left(\beta_j f_E^t(X_i; t^l, t, r_j, \theta)\right) \,\middle|\, t^l \le l_i < u_i \le t < T; \boldsymbol{\Theta}_1^{(h-1)}\right)$$

$$+ \sum_{i \in S_{v\cdot}} E\left(\sum_{j=1}^{M} Z_{ij} \ln\left(\beta_j f_E^t(X_i; t^l, t, r_j, \theta)\right) \,\middle|\, t^l \le l_i < X_i \le t < u_i \le T; \boldsymbol{\Theta}_1^{(h-1)}\right)$$

$$\times P\left(X_i \le t \,\middle|\, t^l \le l_i < t < u_i \le T; \boldsymbol{\Theta}^{(h-1)}\right). \tag{D.3}$$

Uncensored data. Following Verbelen et al. (2015) we obtain that the first part of the sum is equal to

$$\sum_{i \in S_{i\cdot}} \sum_{j=1}^{M} {}^{i\cdot}z_{ij}^{(h)} \Big[\ln \beta_j + (r_j - 1) \ln x_i - \frac{x_i}{\theta} - r_j \ln \theta$$

$$- \ln((r_j - 1)!) - \ln\left(F_E(t; r_j, \theta) - F_E(t^l; r_j, \theta)\right) \Big].$$

Here is

$$
\begin{aligned}
{}^{i.}z_{ij}^{(h)} &:= P\left(Z_{ij}=1 \,\middle|\, t^l \le l_i = u_i \le t < T; \mathbf{\Theta}_1^{(h-1)}\right) \\
&= \frac{\alpha_j^{(h-1)} f_E(x_i; r_j, \theta^{(h-1)})}{\sum_{m=1}^{M} \alpha_m^{(h-1)} f_E(x_i; r_m, \theta^{(h-1)})}
\end{aligned} \tag{D.4}
$$

the posterior probability that an uncensored data point x_i with $x_i \le t$ (hence $i \in S_{i.}$) belongs to the jth component in the mixture. In the E-step for the uncensored data, only these probabilities need to be computed for all $i \in S_{i.}$ and $j = 1, \ldots, M$. They remain the same in the truncated and the untruncated case.

Censored data. We now have to distinguish between the cases *iii* and *v*. The derivations for case *iii* follow from Verbelen et al. (2015).

Denote by ${}^{iii.}z_{ij}^{(h)}$ the posterior probability that the data point x_i belongs to the jth component in the mixture for a censored data point with $u_i \le t$. Then,

$$
\begin{aligned}
&\sum_{i \in S_{iii.}} E\left(\sum_{j=1}^{M} Z_{ij} \ln\left(\beta_j f_E^t(X_i; t^l, t, r_j, \theta)\right) \,\middle|\, t^l \le l_i < u_i \le t < T; \mathbf{\Theta}_1^{(h-1)}\right) \\
&= \sum_{i \in S_{iii.}} \sum_{j=1}^{M} {}^{iii.}z_{ij}^{(h)} E\left(\ln\left(\beta_j f_E^t(X_i; t^l, t, r_j, \theta)\right) \,\middle|\, Z_{ij}=1, t^l \le l_i < u_i \le t < T; \theta^{(h-1)}\right) \\
&= \sum_{i \in S_{iii.}} \sum_{j=1}^{M} {}^{iii.}z_{ij}^{(h)} \Bigg[\ln \beta_j + (r_j - 1) E\left(\ln X_i \,\middle|\, Z_{ij}=1, t^l \le l_i < u_i \le t < T; \theta^{(h-1)}\right) \\
&\qquad - \frac{1}{\theta} E\left(X_i \,\middle|\, Z_{ij}=1, t^l \le l_i < u_i \le t < T; \theta^{(h-1)}\right) - r_j \ln \theta - \ln((r_j-1)!) \\
&\qquad - \ln\left(F_E(t; r_j, \theta) - F_E(t^l; r_j, \theta)\right) \Bigg]
\end{aligned} \tag{D.5}
$$

where we use the law of total expectation in the first equality. Using Bayes' rule, we can compute the posterior probabilities ${}^{iii.}z_{ij}^{(h)}$, for $i \in S_{iii.}$ and $j = 1, \ldots, M$, as

$$
\begin{aligned}
{}^{iii.}z_{ij}^{(h)} &:= P\left(Z_{ij}=1 \,\middle|\, t^l \le l_i < u_i \le t < T; \boldsymbol{\Theta}_1^{(h-1)}\right) \\
&= \frac{\beta_j^{(h-1)}\left(F_E^t(u_i;t^l,t,r_j,\theta^{(h-1)}) - F_E^t(l_i;t^l,t,r_j,\theta^{(h-1)})\right)}{\sum_{m=1}^M \beta_m^{(h-1)}\left(F_E^t(u_i;t^l,t,r_m,\theta^{(h-1)}) - F_E^t(l_i;t^l,t,r_m,\theta^{(h-1)})\right)} \\
&= \frac{\beta_j^{(h-1)}\frac{F_E(u_i;r_j,\theta^{(h-1)})-F_E(l_i;r_j,\theta^{(h-1)})}{F_E(t;r_j,\theta^{(h-1)})-F_E(t^l;r_j,\theta^{(h-1)})}}{\sum_{m=1}^M \beta_m^{(h-1)}\frac{F_E(u_i;r_m,\theta^{(h-1)})-F_E(l_i;r_m,\theta^{(h-1)})}{F_E(t;r_m,\theta^{(h-1)})-F_E(t^l;r_m,\theta^{(h-1)})}} \\
&= \frac{\alpha_j^{(h-1)}\left(F_E(u_i;r_j,\theta^{(h-1)}) - F_E(l_i;r_j,\theta^{(h-1)})\right)}{\sum_{m=1}^M \alpha_m^{(h-1)}\left(F_E(u_i;r_m,\theta^{(h-1)}) - F_E(l_i;r_m,\theta^{(h-1)})\right)}\,. \qquad \text{(D.6)}
\end{aligned}
$$

The expression for the posterior probability in the censored case has the same form as in the uncensored case (D.4), but with the densities replaced by the probabilities to lay between the lower and upper censoring points. The terms in (D.5) containing $(r_j-1)E\left(\ln X_i \,\middle|\, Z_{ij}=1, t^l \le l_i < t < u_i \le T; \theta^{(h-1)}\right)$ do not play a role in the EM algorithm as they do not depend on $\boldsymbol{\beta}$ or θ. We also need to compute the following conditional expected value in the E-step:

$$
\begin{aligned}
&E\left(X_i \,\middle|\, Z_{ij}=1, t^l \le l_i < u_i \le t < T; \theta^{(h-1)}\right) \\
&= \int_{l_i}^{u_i} x \frac{f_E(x;r_j,\theta^{(h-1)})}{F_E(u_i;r_j,\theta^{(h-1)}) - F_E(l_i;r_j,\theta^{(h-1)})}\,dx \\
&= \frac{r_j\theta^{(h-1)}}{F_E(u_i;r_j,\theta^{(h-1)}) - F_E(l_i;r_j,\theta^{(h-1)})} \int_{l_i}^{u_i} \frac{x^{r_j}\exp\left(-x/\theta^{(h-1)}\right)}{\left(\theta^{(h-1)}\right)^{r_j+1} r_j!}\,dx \\
&= \frac{r_j\theta^{(h-1)}\left(F_E(u_i;r_j+1,\theta^{(h-1)}) - F_E(l_i;r_j+1,\theta^{(h-1)})\right)}{F_E(u_i;r_j,\theta^{(h-1)}) - F_E(l_i;r_j,\theta^{(h-1)})}\,, \qquad \text{(D.7)}
\end{aligned}
$$

for $i \in S_{iii.}$ and $j = 1, \ldots, M$, which has a closed-form expression.

We perform similar calculations for the case v. Denote by ${}^{v.}z_{ij}^{(h)}$ the posterior probability that the data point x_i belongs to the jth component in the mixture for a censored data point with $\{X_i \le t\}$ and $u_i > t$. The third part of the sum in (D.3), without the probability, is then given by

$$\sum_{i\in S_{v\cdot}} E\left(\sum_{j=1}^{M} Z_{ij}\ln\left(\beta_j f_E^t(X_i;t^l,t,r_j,\theta)\right)\,\middle|\, t^l\le l_i<X_i\le t<u_i\le T;\boldsymbol{\Theta}_1^{(h-1)}\right)$$

$$=\sum_{i\in S_{v\cdot}}\sum_{j=1}^{M} {}^{v\cdot}z_{ij}^{(h)}\Bigg[\ln\beta_j+(r_j-1)E\left(\ln X_i\,\middle|\, Z_{ij}=1,t^l\le l_i<X_i\le t<u_i\le T;\theta^{(h-1)}\right)$$

$$-\frac{1}{\theta}E\left(X_i\,\middle|\, Z_{ij}=1,t^l\le l_i<X_i\le t<u_i\le T,\theta^{(h-1)}\right)-r_j\ln\theta-\ln((r_j-1)!)$$

$$-\ln\left(F_E(t;r_j,\theta)-F_E(t^l;r_j,\theta)\right)\Bigg]. \tag{D.8}$$

The posterior probabilities ${}^{v\cdot}z_{ij}^{(h)}$, for $i\in S_{v\cdot}$ and $j=1,\ldots,M$, are given by

$$\begin{aligned}{}^{v\cdot}z_{ij}^{(h)} &:= P\left(Z_{ij}=1\,\middle|\, t^l\le l_i<X_i\le t<u_i\le T;\boldsymbol{\Theta}_1^{(h-1)}\right)\\ &=\frac{\beta_j^{(h-1)}\left(F_E^t(t;t^l,t,r_j,\theta^{(h-1)})-F_E^t(l_i;t^l,t,r_j,\theta^{(h-1)})\right)}{\sum_{m=1}^{M}\beta_m^{(h-1)}\left(F_E^t(t;t^l,t,r_m,\theta^{(h-1)})-F_E^t(l_i;t^l,t,r_m,\theta^{(h-1)})\right)}\\ &=\frac{\alpha_j^{(h-1)}\left(F_E(t;r_j,\theta^{(h-1)})-F_E(l_i;r_j,\theta^{(h-1)})\right)}{\sum_{m=1}^{M}\alpha_m^{(h-1)}\left(F_E(t;r_m,\theta^{(h-1)})-F_E(l_i;r_m,\theta^{(h-1)})\right)}.\end{aligned} \tag{D.9}$$

This expression corresponds to (D.6) with t instead of u_i because of the conditioning on the event $\{X_i\le t<u_i\}$. Again, the terms in (D.8) containing $(r_j-1)\,E\left(X_i\,\middle|\, Z_{ij}=1,t^l\le l_i<X_i\le t<u_i\le T;\theta^{(h-1)}\right)$ do not play a role in the EM algorithm as they do not depend on $\boldsymbol{\beta}$ or θ. Also,

$$\begin{aligned}&E\left(X_i\,\middle|\, Z_{ij}=1,t^l\le l_i<X_i\le t<u_i\le T;\theta^{(h-1)}\right)\\ &=\int_{l_i}^{t} x\frac{f_E(x;r_j,\theta^{(h-1)})}{F_E(t;r_j,\theta^{(h-1)})-F_E(l_i;r_j,\theta^{(h-1)})}\,dx\\ &=\frac{r_j\theta^{(h-1)}\left(F_E(t;r_j+1,\theta^{(h-1)})-F_E(l_i;r_j+1,\theta^{(h-1)})\right)}{F_E(t;r_j,\theta^{(h-1)})-F_E(l_i;r_j,\theta^{(h-1)})},\end{aligned} \tag{D.10}$$

for $i\in S_{v\cdot}$ and $j=1,\ldots,M$, which has a closed-form expression. Likewise, (D.10) corresponds to (D.7) with t instead of u_i since we condition on the event $\{X_i\le t<u_i\}$.

Pareto distribution

The relevant part for the Pareto parameter ξ is

$$\begin{aligned}
&\sum_{i\in S_{ii.}} E\left(\ln f_2(X_i;t,T,\xi)\,\middle|\, t^l < t < l_i = u_i \le T;\xi^{(h-1)}\right)\\
&+\sum_{i\in S_{iv.}} E\left(\ln f_2(X_i;t,T,\xi)\,\middle|\, t^l < t \le l_i < u_i \le T;\xi^{(h-1)}\right)\\
&+\sum_{i\in S_{v.}} E\left(\ln f_2(X_i;t,T,\xi)\,\middle|\, t^l \le l_i \le t < X_i < u_i \le T;\xi^{(h-1)}\right)\\
&\qquad\times P\left(X_i > t\,\middle|\, t^l \le l_i < t < u_i \le T;\boldsymbol{\Theta}^{(h-1)}\right).
\end{aligned}$$

The first conditional expectation is simply equal to

$$\begin{aligned}
&E\left(\ln f_2(X_i;t,T,\xi)\,\middle|\, t^l < t < l_i = u_i \le T;\xi^{(h-1)}\right)\\
&= \ln f_2(x_i;t,T,\xi) = -\ln(\xi t) - \left(\frac{1}{\xi}+1\right)\ln\left(\frac{x_i}{t}\right) - \ln\left(1-\left(\frac{T}{t}\right)^{-\frac{1}{\xi}}\right).
\end{aligned}$$

Starting from the definition we obtain

$$\begin{aligned}
&E\left(\ln f_2(X_i;t,T,\xi)\,\middle|\, t^l < t \le l_i < u_i \le T;\xi^{(h-1)}\right)\\
&= -\ln(\xi t) - \left(\frac{1}{\xi}+1\right) E\left(\ln\left(\frac{X_i}{t}\right)\,\middle|\, t^l < t \le l_i < u_i \le T;\xi^{(h-1)}\right)\\
&\quad - \ln\left(1-\left(\frac{T}{t}\right)^{-\frac{1}{\xi}}\right).
\end{aligned}$$

Using integration by parts, we get

$$\begin{aligned}
&E\left(\ln\left(\frac{X_i}{t}\right)\,\middle|\, t^l < t \le l_i < u_i \le T;\xi^{(h-1)}\right)\\
&= \int_{l_i}^{u_i} \ln\left(\frac{x}{t}\right)\frac{f_2^*\left(x;t,\xi^{(h-1)}\right)}{F_2^*\left(u_i;t,\xi^{(h-1)}\right)-F_2^*\left(l_i;t,\xi^{(h-1)}\right)}\,dx\\
&= \frac{1}{\left(\frac{l_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}}-\left(\frac{u_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}}}\int_{l_i}^{u_i}\frac{\ln\left(\frac{x}{t}\right)}{\xi^{(h-1)}t}\left(\frac{x}{t}\right)^{-\frac{1}{\xi^{(h-1)}}-1}\,dx
\end{aligned}$$

$$= \frac{1}{\left(\frac{l_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}} - \left(\frac{u_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}}} \int_{\frac{l_i}{t}}^{\frac{u_i}{t}} \frac{\ln v}{\xi^{(h-1)}} v^{-\frac{1}{\xi^{(h-1)}}-1} dv$$

$$= \frac{1}{\left(\frac{l_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}} - \left(\frac{u_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}}} \left(- \left[\ln v \; v^{-\frac{1}{\xi^{(h-1)}}}\right]_{\frac{l_i}{t}}^{\frac{u_i}{t}} + \int_{\frac{l_i}{t}}^{\frac{u_i}{t}} v^{-\frac{1}{\xi^{(h-1)}}-1} dv \right)$$

$$= \frac{1}{\left(\frac{l_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}} - \left(\frac{u_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}}} \left(- \left[\ln v \; v^{-\frac{1}{\xi^{(h-1)}}}\right]_{\frac{l_i}{t}}^{\frac{u_i}{t}} + \left[-\xi^{(h-1)} v^{-\frac{1}{\xi^{(h-1)}}}\right]_{\frac{l_i}{t}}^{\frac{u_i}{t}} \right)$$

$$= \frac{\left(\ln\left(\frac{l_i}{t}\right) + \xi^{(h-1)}\right)\left(\frac{l_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}} - \left(\ln\left(\frac{u_i}{t}\right) + \xi^{(h-1)}\right)\left(\frac{u_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}}}{\left(\frac{l_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}} - \left(\frac{u_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}}}. \tag{D.11}$$

We compute the third conditional expectation similarly which gives

$$E\Big(\ln f_2(X_i; t, T, \xi) \,\Big|\, t^l \le l_i \le t < X_i < u_i \le T; \xi^{(h-1)} \Big)$$

$$= -\ln(\xi t) - \left(\frac{1}{\xi} + 1\right) E\left(\ln\left(\frac{X_i}{t}\right) \,\middle|\, t^l \le l_i \le t < X_i < u_i \le T; \xi^{(h-1)} \right)$$

$$- \ln\left(1 - \left(\frac{T}{t}\right)^{-\frac{1}{\xi}}\right),$$

with

$$E\left(\ln\left(\frac{X_i}{t}\right) \,\middle|\, t^l \le l_i \le t < X_i < u_i \le T; \xi^{(h-1)} \right) =$$

$$\frac{\xi^{(h-1)} - \left(\ln\left(\frac{u_i}{t}\right) + \xi^{(h-1)}\right)\left(\frac{u_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}}}{1 - \left(\frac{u_i}{t}\right)^{-\frac{1}{\xi^{(h-1)}}}}. \tag{D.12}$$

Note that the last expression corresponds to (D.11) with $l_i = t$ since we condition on the event $\{l_i \le t < X_i\}$.

D.1.3 M-step

In the M-step, the expected value of the complete data log-likelihood obtained in the E-step is maximised w.r.t. the parameter vector $\boldsymbol{\Theta}$. As said before, we assume that there are no shared parameters in π, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$. This assures that we can split the maximisation in three parts.

Splicing weight π

Maximisation with respect to π gives

$$\pi^{(h)} = \frac{n_1^{(h)}}{n} = \frac{\#S_{i\cdot} + \#S_{iii\cdot} + \sum_{i\in S_{v\cdot}} P\left(X_i \le t \,\middle|\, t^l \le l_i < t < u_i \le T; \boldsymbol{\Theta}^{(h-1)}\right)}{n},$$

see (6.10). An expression for the probability can be found in (D.2).

ME distribution

The expected value of the complete data log-likelihood obtained in the E-step is now maximised with respect to the parameter vector $\boldsymbol{\Theta}_1 = (\boldsymbol{\beta}, \theta)$ over all $(\boldsymbol{\beta}, \theta)$ with $\beta_j > 0$, $\sum_{j=1}^M \beta_j = 1$ and $\theta > 0$.

To maximise over the mixing weights $\boldsymbol{\beta}$,

$$\sum_{j=1}^{M} \left(\sum_{i\in S_{i\cdot}} {}^{i\cdot}z_{ij}^{(h)} + \sum_{i\in S_{iii\cdot}} {}^{iii\cdot}z_{ij}^{(h)} + \sum_{i\in S_{v\cdot}} P_{1,i}^{(h)}\, {}^{v\cdot}z_{ij}^{(h)} \right) \ln \beta_j$$

needs to be maximised where $P_{1,i}^{(h)} = P\left(X_i \le t \,\middle|\, t^l \le l_i < t < u_i \le T; \boldsymbol{\Theta}^{(h-1)}\right)$, see (D.2). Setting $\beta_M = 1 - \sum_{j=1}^{M-1} \beta_j$ makes sure that the restriction $\sum_{j=1}^M \beta_j = 1$ holds. Equating the partial derivatives at $\boldsymbol{\beta}^{(h)}$ to zero gives

$$\beta_j^{(h)} = \frac{\sum_{i\in S_{i\cdot}} {}^{i\cdot}z_{ij}^{(h)} + \sum_{i\in S_{iii\cdot}} {}^{iii\cdot}z_{ij}^{(h)} + \sum_{i\in S_{v\cdot}} P_{1,i}^{(h)}\, {}^{v\cdot}z_{ij}^{(h)}}{\sum_{i\in S_{i\cdot}} {}^{i\cdot}z_{iM}^{(h)} + \sum_{i\in S_{iii\cdot}} {}^{iii\cdot}z_{iM}^{(h)} + \sum_{i\in S_{v\cdot}} P_{1,i}^{(h)}\, {}^{v\cdot}z_{iM}^{(h)}} \beta_M^{(h)}$$

for $j = 1, \ldots, M-1$. Applying the sum constraint gives

$$\beta_M^{(h)} = \frac{\sum_{i\in S_{i\cdot}} {}^{i\cdot}z_{iM}^{(h)} + \sum_{i\in S_{iii\cdot}} {}^{iii\cdot}z_{iM}^{(h)} + \sum_{i\in S_{v\cdot}} P_{1,i}^{(h)}\, {}^{v\cdot}z_{iM}^{(h)}}{n_1^{(h)}}.$$

The same form also follows for $j = 1, \ldots, M-1$:

$$\beta_j^{(h)} = \frac{\sum_{i\in S_{i\cdot}} {}^{i\cdot}z_{ij}^{(h)} + \sum_{i\in S_{iii\cdot}} {}^{iii\cdot}z_{ij}^{(h)} + \sum_{i\in S_{v\cdot}} P_{1,i}^{(h)}\, {}^{v\cdot}z_{ij}^{(h)}}{n_1^{(h)}}. \tag{D.13}$$

The estimate for the prior probability β_j in the truncated mixture is thus the average of the posterior probabilities of belonging to the jth component in the mixture.

In order to maximise with respect to θ, we set the first order partial derivative at $\theta^{(h)}$ equal to zero

$$
\begin{aligned}
0 = \sum_{j=1}^{M} \Bigg(& \sum_{i \in S_{i.}} {}^{i.}z_{ij}^{(h)} \left(-\frac{r_j}{\theta} - \frac{\frac{\partial}{\partial\theta}\left[F_E(t;r_j,\theta) - F_E(t^l;r_j,\theta)\right]}{F_E(t;r_j,\theta) - F_E(t^l;r_j,\theta)} + \frac{x_i}{\theta^2} \right) \\
& + \sum_{i \in S_{iii.}} {}^{iii.}z_{ij}^{(h)} \left(-\frac{r_j}{\theta} - \frac{\frac{\partial}{\partial\theta}\left[F_E(t;r_j,\theta) - F_E(t^l;r_j,\theta)\right]}{F_E(t;r_j,\theta) - F_E(t^l;r_j,\theta)} \right. \\
& \qquad \left. + \frac{E\left(X_i \,\middle|\, Z_{ij}=1, t^l \le l_i < u_i \le t < T; \theta^{(h-1)}\right)}{\theta^2} \right) \\
& + \sum_{i \in S_{v.}} P_{1,i}^{(h)\ v.}z_{ij}^{(h)} \left(-\frac{r_j}{\theta} - \frac{\frac{\partial}{\partial\theta}\left[F_E(t;r_j,\theta) - F_E(t^l;r_j,\theta)\right]}{F_E(t;r_j,\theta) - F_E(t^l;r_j,\theta)} \right. \\
& \qquad \left. \left. + \frac{E\left(X_i \,\middle|\, Z_{ij}=1, t^l \le l_i < X_i \le t < u_i \le T; \theta^{(h-1)}\right)}{\theta^2} \right) \right) \Bigg|_{\theta=\theta^{(h)}} \\
= & -\frac{1}{\theta^{(h)}} \sum_{j=1}^{M} \left(\sum_{i \in S_{i.}} {}^{i.}z_{ij}^{(h)} + \sum_{i \in S_{iii.}} {}^{iii.}z_{ij}^{(h)} + \sum_{i \in S_{v.}} P_{1,i}^{(h)\ v.}z_{ij}^{(h)} \right) r_j \\
& - \sum_{j=1}^{M} \left(\sum_{i \in S_{i.}} {}^{i.}z_{ij}^{(h)} + \sum_{i \in S_{iii.}} {}^{iii.}z_{ij}^{(h)} + \sum_{i \in S_{v.}} P_{1,i}^{(h)\ v.}z_{ij}^{(h)} \right) \\
& \qquad \times \left. \frac{\frac{\partial}{\partial\theta}\left[F_E(t;r_j,\theta) - F_E(t^l;r_j,\theta)\right]}{F_E(t;r_j,\theta) - F_E(t^l;r_j,\theta)} \right|_{\theta=\theta^{(h)}} \\
& + \frac{1}{\theta^{(h)2}} \sum_{j=1}^{M} \Bigg(\sum_{i \in S_{i.}} {}^{i.}z_{ij}^{(h)}\ x_i \\
& \qquad + \sum_{i \in S_{iii.}} {}^{iii.}z_{ij}^{(h)} E\left(X_i \,\middle|\, Z_{ij}=1, t^l \le l_i < u_i \le t < T; \theta^{(h-1)}\right) \\
& \qquad + \sum_{i \in S_{v.}} P_{1,i}^{(h)\ v.}z_{ij}^{(h)} E\left(X_i \,\middle|\, Z_{ij}=1, t^l \le l_i < X_i \le t < u_i \le T; \theta^{(h-1)}\right) \Bigg).
\end{aligned}
$$

Using the lower incomplete gamma function

$$\gamma(s,x) = \int_0^x z^{s-1}\exp(-z)\,dz\,,$$

we write the cumulative distribution function of an Erlang distribution as

$$F_E(x;r_j,\theta) = \int_0^x \frac{z^{r-1}\exp(-z/\theta)}{\theta^{r_j}(r_j-1)!}\,dz = \frac{1}{(r-1)!}\int_0^{x/\theta} u^{r_j-1}\exp(-u)\,du = \frac{\gamma(r_j,x/\theta)}{(r_j-1)!}\,.$$

Applying the Leibniz rule gives that the partial derivative of F_E with respect to θ is equal to

$$\frac{\partial F_E(x;r_j,\theta)}{\partial\theta} = \frac{\frac{\partial\gamma(r_j,x/\theta)}{\partial\theta}}{(r_j-1)!} = \frac{\left(\frac{x}{\theta}\right)^{r_j-1}\exp\left(-\frac{x}{\theta}\right)\left(-\frac{x}{\theta^2}\right)}{(r_j-1)!}.$$

Using (D.13) and this derivative gives

$$\begin{aligned}
0 = &-\frac{n_1^{(h)}}{\theta^{(h)}}\sum_{j=1}^{M}\beta_j^{(h)}r_j \\
&- n_1^{(h)}\sum_{j=1}^{M}\beta_j^{(h)}\left.\frac{\left(\frac{t^l}{\theta}\right)^{r_j-1}\exp\left(-\frac{t^l}{\theta}\right)\frac{t^l}{\theta^2} - \left(\frac{t}{\theta}\right)^{r_j-1}\exp\left(-\frac{t}{\theta}\right)\frac{t}{\theta^2}}{(r_j-1)!\,(F_E(t;r_j,\theta)-F_E(t^l;r_j,\theta))}\right|_{\theta=\theta^{(h)}} \\
&+ \frac{1}{\theta^{(h)2}}\sum_{j=1}^{M}\Bigg(\sum_{i\in S_{i\cdot}}{}^{i\cdot}z_{ij}^{(h)}\;x_i \\
&\qquad + \sum_{i\in S_{iii\cdot}}{}^{iii\cdot}z_{ij}^{(h)}E\left(X_i\,\middle|\,Z_{ij}=1,t^l\le l_i<u_i\le t<T;\theta^{(h-1)}\right) \\
&\qquad + \sum_{i\in S_{v\cdot}}P_{1,i}^{(h)}\;{}^{v\cdot}z_{ij}^{(h)}E\left(X_i\,\middle|\,Z_{ij}=1,t^l\le l_i<X_i\le t<u_i\le T;\theta^{(h-1)}\right)\Bigg).
\end{aligned}$$

This leads to the following M-step equation for θ:

$$\theta^{(h)} = \frac{1}{n_1^{(h)} \sum_{j=1}^{M} \beta_j^{(h)} r_j} \Bigg(\sum_{i \in S_{i\cdot}} x_i + \sum_{i \in S_{iii\cdot}} \sum_{j=1}^{M} {}^{iii\cdot} z_{ij}^{(h)} E\left(X_i \,\middle|\, Z_{ij} = 1, t^l \leq l_i < u_i \leq t < T; \theta^{(h-1)} \right) + \sum_{i \in S_{v\cdot}} \sum_{j=1}^{M} P_{1,i}^{(h)\ v\cdot} z_{ij}^{(h)} E\left(X_i \,\middle|\, Z_{ij} = 1, t^l \leq l_i < X_i \leq t < u_i \leq T; \theta^{(h-1)} \right) \Bigg)$$

$$- \frac{\sum_{j=1}^{M} \beta_j^{(h)} \frac{\left(t^l\right)^{r_j} \exp\left(-\frac{t^l}{\theta^{(h)}}\right) - t^{r_j} \exp\left(-\frac{t}{\theta^{(h)}}\right)}{\left(\theta^{(h)}\right)^{r_j - 1} (r_j - 1)! \left(F_E(t; r_j, \theta^{(h)}) - F_E(t^l; r_j, \theta^{(h)})\right)}}{\sum_{j=1}^{M} \beta_j^{(h)} r_j}, \tag{D.14}$$

where expressions for the conditional expectations can be found in (D.7) and (D.10). This equation can be seen as the sample mean of all data points that are smaller than or equal to t, divided by the average shape parameter. For the censored data points, the sample mean is replaced by the expected value given the lower and upper bounds (second and third terms). Moreover, there is a correction for truncation (fourth term) which depends on $\theta^{(h)}$ itself in a complicated way. Therefore, it is not possible to find an analytical solution, and we solve (D.14) numerically using a Newton-type algorithm with the previous estimate $\theta^{(h-1)}$ as the starting value.

The expression for β_j and the M-step equation for θ are similar to the ones in Verbelen et al. (2015) where an extra term (third term) is included for data points of type v.

Then, we transform the estimates for β_j to estimates for α_j using $\widetilde{\alpha}_j / \sum_{j=1}^{M} \widetilde{\alpha}_j$ with

$$\widetilde{\alpha}_j = \frac{\hat{\beta}_j}{F_E(t; r_j, \hat{\theta}) - F_E(t^l; r_j, \hat{\theta})},$$

and $\hat{\beta}_j$ and $\hat{\theta}$ the estimates obtained from the final EM-step.

Pareto distribution

To maximise the expected log-likelihood with respect to $\mathbf{\Theta}_2 = \xi$, we have to maximise

$$
\begin{aligned}
&\sum_{i \in S_{ii.}} E\left(\ln f_2(X_i;t,T,\xi) \,\middle|\, t^l < t < l_i = u_i \le T; \xi^{(h-1)}\right) \\
&+ \sum_{i \in S_{iv.}} E\left(\ln f_2(X_i;t,T,\xi) \,\middle|\, t^l < t \le l_i < u_i \le T; \xi^{(h-1)}\right) \\
&+ \sum_{i \in S_{v.}} P_{2,i}^{(h)} E\left(\ln f_2(X_i;t,T,\xi) \,\middle|\, t^l \le l_i \le t < X_i < u_i \le T; \xi^{(h-1)}\right) \\
&= \sum_{i \in S_{ii.}} \Big[-\ln(\xi t) - \ln\left(1 - \left(\frac{T}{t}\right)^{-\frac{1}{\xi}}\right) - \left(\frac{1}{\xi}+1\right)\ln\left(\frac{x_i}{t}\right) \Big] \\
&+ \sum_{i \in S_{iv.}} \Bigg[-\ln(\xi t) - \ln\left(1 - \left(\frac{T}{t}\right)^{-\frac{1}{\xi}}\right) \\
&\qquad - \left(\frac{1}{\xi}+1\right) E\left(\ln\left(\frac{X_i}{t}\right) \,\middle|\, t^l < t \le l_i < u_i \le T; \xi^{(h-1)}\right) \Bigg] \\
&+ \sum_{i \in S_{v.}} P_{2,i}^{(h)} \Bigg[-\ln(\xi t) - \ln\left(1 - \left(\frac{T}{t}\right)^{-\frac{1}{\xi}}\right) \\
&\qquad - \left(\frac{1}{\xi}+1\right) E\left(\ln\left(\frac{X_i}{t}\right) \,\middle|\, t^l \le l_i \le t < X_i < u_i \le T; \xi^{(h-1)}\right) \Bigg],
\end{aligned}
\tag{D.15}
$$

where $P_{2,i}^{(h)} = P\left(X_i > t \,\middle|\, t^l \le l_i < t < u_i \le T; \mathbf{\Theta}^{(h-1)}\right)$. Taking the derivative of (D.15) with respect to ξ gives

$$
\begin{aligned}
&-\frac{1}{\xi}\left(\#S_{ii.} + \#S_{iv.} + \sum_{i \in S_{v.}} P_{2,i}^{(h)}\right) \\
&+ \frac{1}{\xi^2}\left(\#S_{ii.} + \#S_{iv.} + \sum_{i \in S_{v.}} P_{2,i}^{(h)}\right) \frac{\ln\left(\frac{T}{t}\right)\left(\frac{T}{t}\right)^{-\frac{1}{\xi}}}{1 - \left(\frac{T}{t}\right)^{-\frac{1}{\xi}}}
\end{aligned}
$$

$$+\frac{1}{\xi^2}\Bigg(\sum_{i\in S_{ii.}}\ln\left(\frac{x_i}{t}\right)+\sum_{i\in S_{iv.}}E\left(\ln\left(\frac{X_i}{t}\right)\middle|\, t^l<t\le l_i<u_i\le T;\xi^{(h-1)}\right)$$

$$+\sum_{i\in S_{v.}}P_{2,i}^{(h)}\;E\left(\ln\left(\frac{X_i}{t}\right)\middle|\, t^l\le l_i\le t<X_i<u_i\le T;\xi^{(h-1)}\right)\Bigg).$$

Setting this derivative at $\xi^{(h)}$ equal to 0 and then solving for $\xi^{(h)}$ results in

$$\begin{aligned}\xi^{(h)}=&\frac{1}{n_2^{(h)}}\Bigg(\sum_{i\in S_{ii.}}\ln\left(\frac{x_i}{t}\right)\\&+\sum_{i\in S_{iv.}}E\left(\ln\left(\frac{X_i}{t}\right)\middle|\, t^l<t\le l_i<u_i\le T;\xi^{(h-1)}\right)\\&+\sum_{i\in S_{v.}}P_{2,i}^{(h)}\;E\left(\ln\left(\frac{X_i}{t}\right)\middle|\, t^l\le l_i\le t<X_i<u_i\le T;\xi^{(h-1)}\right)\Bigg)\\&+\frac{\ln\left(\frac{T}{t}\right)}{\left(\frac{T}{t}\right)^{\frac{1}{\xi^{(h)}}}-1}.\end{aligned}$$

where expressions for the conditional expectations can be found in (D.11) and (D.12). Note that $P_{2,i}^{(h)}=1-P_{1,i}^{(h)}$ and we compute the latter using (D.2). The first term in the equation can be seen as the sample mean of all $\ln(X_i/t)$ with $\{X_i>t\}$ which are observed for data points belonging to case *ii* and have to be replaced by their expected values for cases *iv* and *v*. Similarly to (D.14), the last term, which depends on $\xi^{(h)}$, is a correction term for the upper truncation at point T and requires us to solve this equation numerically. In case there is no upper truncation, i.e. $T=+\infty$, the last term is equal to zero and we obtain an analytical solution for $\xi^{(h)}$.

D.1.4 Choice of shape parameters and number of mixtures for ME distribution

To choose the shape parameters $\boldsymbol{r}$ and the number of mixtures M, we follow the approach from Verbelen et al. (2016) which is a slightly modified version of the approach from Verbelen et al. (2015). Starting from a certain value for M and shapes as given in (D.1), they try to reduce M using a backward stepwise search where the mixture component with the smallest shape is deleted if this decreases an IC (AIC or BIC). Moreover, after the first reduction of M, M is

further reduced using the IC and the shapes $\boldsymbol{r}$ are adjusted based on maximising the likelihood. For each value for M and $\boldsymbol{r}$, the EM algorithm described above is executed. We refer to Section 4 in Verbelen et al. (2016) for more details. Important is that we now consider the (log-)likelihood including the Pareto part, as given in (6.6).

D.2 Fitting the ME-Pareto model to uncensored data using the EM algorithm

In case there are no censored data points, calculations are much simpler. Now, we only have data points from cases i and ii.

D.2.1 Starting values

We use the starting values for θ and $\boldsymbol{\beta}$ from Verbelen et al. (2016), see Section D.1.1. Since the obtained estimates are the same in each step, we do not need starting values for π and ξ.

D.2.2 Splicing weight π

Maximisation w.r.t. π gives

$$\pi^{(h)} = \frac{\#S_{i\cdot}}{n} = \frac{n_1}{n}. \tag{D.16}$$

The estimate for π is thus equal to the proportion of points that is smaller than or equal to the splicing point t. It is clear that $\pi^{(h)}$ is constant since it is independent of h.

Taking the splicing point equal to the $(k+1)$th largest data point, i.e. $t = x_{n-k,n}$, gives

$$\pi^{(h)} = 1 - \frac{k}{n}$$

when there are no ties.

D.2.3 ME distribution

Similar to before, the following expression is obtained, for $j = 1, \ldots, M$,

$$\beta_j^{(h)} = \frac{\sum_{i \in S_{i\cdot}} {}^{i\cdot} z_{ij}^{(h)}}{n_1}.$$

The M-step equation for θ now simplifies to

$$\theta^{(h)} = \frac{\sum_{i \in S_{i\cdot}} x_i}{n_1 \sum_{j=1}^{M} \beta_j^{(h)} r_j} - \frac{\sum_{j=1}^{M} \beta_j^{(h)} \frac{\left(t^l\right)^{r_j} \exp\left(-\frac{t^l}{\theta^{(h)}}\right) - t^{r_j} \exp\left(-\frac{t}{\theta^{(h)}}\right)}{\left(\theta^{(h)}\right)^{r_j - 1} (r_j - 1)! \left(F_E(t; r_j, \theta^{(h)}) - F_E(t^l; r_j, \theta^{(h)})\right)}}{\sum_{j=1}^{M} \beta_j^{(h)} r_j}. \tag{D.17}$$

As before, a Newton-type algorithm is employed to solve (D.17) numerically using the previous estimate $\theta^{(h-1)}$ as starting value.

D.2.4 Pareto distribution

For the Pareto parameter ξ we get following implicit equation

$$\xi^{(h)} = \frac{\sum_{i \in S_{ii\cdot}} \ln\left(\frac{x_i}{t}\right)}{n_2} + \frac{\ln\left(\frac{T}{t}\right)}{\left(\frac{T}{t}\right)^{\frac{1}{\xi^{(h)}}} - 1}. \tag{D.18}$$

It is immediately clear that this estimate does not depend on h, so we denote $\hat{\xi} = \xi^{(h)}$. The first term of (D.18) is the Hill estimator (Hill, 1975) with threshold t, and the second term is a correction for the upper truncation at T which vanishes for $T = +\infty$. Setting t equal to the $(k+1)$th largest data point $x_{n-k,n}$ gives

$$\hat{\xi} = \frac{1}{k} \sum_{i=1}^{k} \ln\left(\frac{x_{n-i+1,n}}{x_{n-k,n}}\right) + \frac{\ln\left(\frac{T}{x_{n-k,n}}\right)}{\left(\frac{T}{x_{n-k,n}}\right)^{\frac{1}{\hat{\xi}}} - 1} = H_{k,n} + \frac{\ln\left(\frac{T}{x_{n-k,n}}\right)}{\left(\frac{T}{x_{n-k,n}}\right)^{\frac{1}{\hat{\xi}}} - 1}. \tag{D.19}$$

Note that (4.21) corresponds to (D.19) where T is estimated by its conditional MLE $x_{n,n}$ (see also Section 6.4.3).

Bibliography

Aban, I. B., Meerschaert, M. M. and Panorska, A. K. (2006). Parameter Estimation for the Truncated Pareto Distribution. *J. Amer. Statist. Assoc.*, **101**(473), 270–277.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Trans. Automat. Control*, **19**(6), 716–723.

Aki, K. (1965). Maximum Likelihood Estimate of b in the Formula $\log(N) = a - bM$ and its Confidence Limits. *Bull. Earthq. Res. Inst. Tokyo Univ.*, **43**(2), 237–239.

Akritas, M. G. and Van Keilegom, I. (2003). Estimation of Bivariate and Marginal Distributions With Censored Data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **65**(2), 457–471.

Albrecher, H., Beirlant, J. and Teugels, J. (2017). *Reinsurance: Actuarial and Statistical Aspects*. John Wiley & Sons, Ltd, Chichester, UK. To appear.

Antonio, K. and Plat, R. (2014). Micro-Level Stochastic Loss Reserving for General Insurance. *Scand. Actuar. J.*, **2014**(7), 649–669.

Aue, F. and Kalkbrener, M. (2006). LDA at Work: Deutsche Bank's Approach to Quantifying Operational Risk. *J. Oper. Risk*, **1**(4), 49–93.

Babu, G. J. and Rao, C. R. (2004). Goodness-of-fit Tests When Parameters are Estimated. *Sankhyā: The Indian Journal of Statistics*, **66**(1), 63–74.

Bakar, S. A. A., Hamzah, N. A., Maghsoudi, M. and Nadarajah, S. (2015). Modeling Loss Data Using Composite Models. *Insurance Math. Econom.*, **61**, 1146–1154.

Balkema, A. A. and de Haan, L. (1974). Residual Life Time at Great Age. *Ann. Probab.*, **2**(5), 792–804.

Beirlant, J., Fraga Alves, I. and Reynkens, T. (2017). Fitting Tails Affected by Truncation. *Electron. J. Stat.*, **11**(1), 2026–2065.

Beirlant, J., Fraga Alves, M. I. and Gomes, M. I. (2016a). Tail Fitting for Truncated and Non-Truncated Pareto-Type Distributions. *Extremes*, **19**(3), 429–462.

Beirlant, J., Goegebeur, Y., Teugels, J. and Segers, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley, Chichester.

Beirlant, J., Guillou, A., Dierckx, G. and Fils-Villetard, A. (2007). Estimation of the Extreme Value Index and Extreme Quantiles Under Random Censoring. *Extremes*, **10**(3), 151–174.

Beirlant, J., Joossens, E. and Segers, J. (2009). Second-Order Refined Peaks-Over-Threshold Modelling for Heavy-Tailed Distributions. *J. Statist. Plann. Inference*, **139**(8), 2800–2815.

Beirlant, J., Schoutens, W., De Spiegeleer, J., Reynkens, T. and Herrmann, K. (2016b). Hunting for Black Swans in the European Banking Sector Using Extreme Value Analysis. In: J. Kallsen and A. Papapantoleon (eds.), *Advanced Modelling in Mathematical Finance: In Honour of Ernst Eberlein*, Springer International Publishing, Switzerland, pp. 147–166.

Beirlant, J., Schoutens, W. and Segers, J. (2005). Mandelbrot's Extremism. *Wilmott Magazine*, 97–103.

Beirlant, J., Vynckier, P. and Teugels, J. L. (1996). Tail Index Estimation, Pareto Quantile Plots and Regression Diagnostics. *J. Amer. Statist. Assoc.*, **91**(436), 1659–1667.

Björck, Å. and Golub, G. H. (1973). Numerical Methods for Computing Angles Between Linear Subspaces. *Math. Comp.*, **27**(123), 579–594.

Bollerslev, T. and Todorov, V. (2011). Tails, Fears and Risk Premia. *J. Finance*, **66**(6), 2165–2211.

Bourne, S. J., Oates, S. J., van Elk, J. and Doornhof, D. (2014). A Seismological Model for Earthquakes Induced by Fluid Extraction From a Subsurface Reservoir. *J. Geophys. Res. Solid Earth*, **119**(12), 8991–9015.

Brockett, P. L., Derrig, R. A., Golden, L. A., Levine, A. and Alpert, M. (2002). Fraud Classification Using Principal Component Analysis of RIDITs. *J. Risk Insur.*, **69**(3), 341–371.

Bross, I. D. J. (1958). How to Use Ridit Analysis. *Biometrics*, **14**(1), 18–38.

Brys, G., Hubert, M. and Rousseeuw, P. J. (2005). A Robustification of Independent Component Analysis. *J. Chemometr.*, **19**(5-7), 364–375.

Brys, G., Hubert, M. and Struyf, A. (2004). A Robust Measure of Skewness. *J. Comput. Graph. Statist.*, **13**(4), 996–1017.

Cadima, J. and Jolliffe, I. T. (1995). Loadings and Correlations in the Interpretation of Principal Components. *J. Appl. Stat.*, **22**(2), 203–214.

Calderín–Ojeda, E. and Kwok, C. F. (2016). Modeling Claims Data With Composite Stoppa Models. *Scand. Actuar. J.*, **2016**(9), 817–836.

Candès, E. J., Li, X., Ma, Y. and Wright, J. (2011). Robust Principal Component Analysis? *J. ACM*, **58**(3), 1–37.

Cao, R., Vilar, J. M. and Devía, A. (2009). Modelling Consumer Credit Risk via Survival Analysis. *SORT*, **33**(1), 3–30.

Castillo, E., Hadi, A., Balakrishnan, N. and Sarabia, J. (2005). *Extreme Value and Related Models With Applications in Engineering and Science*. Wiley, Hoboken, NJ.

Chakrabarty, A. and Samorodnitsky, G. (2012). Understanding Heavy Tails in a Bounded World, or, is a Truncated Heavy Tail Heavy or not? *Stoch. Models*, **28**(1), 109–143.

Ciumara, R. (2006). An Actuarial Model Based on the Composite Weibull-Pareto Distribution. *Math. Rep. (Bucur.),* **8**(4), 401–414.

Coles, S. (2001). *An Introduction to Statistical Modelling of Extreme Values*. Springer-Verlag, London.

Cooke, P. (1979). Statistical Inference for Bounds of Random Variables. *Biometrika*, **66**(2), 367–374.

Cooke, P. (1980). Optimal Linear Estimation of Bounds of Random Variables. *Biometrika*, **67**(1), 257–258.

Cooray, K. and Ananda, M. M. (2005). Modeling Actuarial Data With a Composite Lognormal-Pareto Model. *Scand. Actuar. J.*, **2005**(5), 321–334.

Cornell, C. A. (1994). Statistical Analysis of Maximum Magnitudes. In: J. F. Schneider (eds.), *The Earthquakes of Stable Continental Regions. Vol. 1: Assessment of Large Earthquake Potential*, EPRI, Palo Alto, CA, pp. 5-1–5-27.

Croux, C., Filzmoser, P. and Fritz, H. (2013). Robust Sparse Principal Component Analysis. *Technometrics*, **55**(2), 202–214.

Croux, C., Filzmoser, P. and Oliveira, M. R. (2007). Algorithms for Projection-Pursuit Robust Principal Component Analysis. *Chemometr. Intell. Lab. Syst.*, **87**(2), 218–225.

Croux, C. and Ruiz-Gazen, A. (2005). High Breakdown Estimators for Principal Components: the Projection-Pursuit Approach Revisited. *J. Multivariate Anal.*, **95**(1), 206–226.

Csörgő, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley, Chichester.

Davies, N. and Kijko, A. (2003). Seismic Risk Assessment: With an Application to the South African Insurance Industry. *S. Afr. Actuar. J.*, **3**(1), 1–28.

de Haan, L. (1970). *On Regular Variation and its Applications to the Weak Convergence of Sample Extremes*. Mathematical Centre Tract 32, Amsterdam.

de Haan, L. (1984). Slow Variation and Characterization of Domains of Attraction. In: T. de Oliveira (eds.), *Statistical Extremes and Applications*, D. Reidel, Dordrecht, pp. 31–48.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: an Introduction*. Springer-Verlag, New York, NY.

Debruyne, M. and Hubert, M. (2009). The Influence Function of the Stahel-Donoho Covariance Estimator of Smallest Outlyingness. *Statist. Probab. Lett.*, **79**(3), 275–282.

Dekkers, A. L. M., Einmahl, J. H. J. and de Haan, L. (1989). A Moment Estimator for the Index of an Extreme-Value Distribution. *Ann. Statist.*, **17**(4), 1795–1832.

Dell'Aquila, R. and Embrechts, P. (2006). Extremes and Robustness: A Contradiction? *Fin. Mkts. Portfolio Mgmt.*, **20**(1), 103–118.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood From Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **39**(1), 1–38.

Dierckx, G. and Teugels, J. L. (2010). Change Point Analysis of Extreme Values. *Environmetrics*, **21**(7-8), 661–686.

Dost, B., Caccavale, M., van Eck, T. and Kraaijpoel, D. (2013). *Report on the Expected PGV and PGA Values for Induced Earthquakes in the Groningen Area*. URL: `https://www.rijksoverheid.nl/documenten/rapporten/2014/01/17/rapport-verwachte-maximale-magnitude-van-aardbevingen-in-groningen`, KNMI report; last accessed on 21/02/2017.

Dost, B. and Kraaijpoel, D. (2013). *The August 16, 2012 Earthquake Near Huizinge (Groningen)*. URL: `https://www.rijksoverheid.nl/documenten/rapporten/2013/01/15/the-august-16-2012-earthquake-near-huizinge-groningen`, KNMI report; last accessed on 25/04/2017.

Drees, H. and Müller, P. (2008). Fitting and Validation of a Bivariate Model for Large Claims. *Insurance Math. Econom.*, **42**(2), 638–650.

Dupuis, D. J. and Field, C. A. (1998). Robust Estimation of Extremes. *Canad. J. Statist*, **26**(2), 199–215.

Dutang, C., Goulet, V. and Pigeon, M. (2008). actuar: An `R` package for Actuarial Science. *J. Stat. Softw.*, **25**(7), 1–37.

Dutang, C. and Jaunatre, K. (2017). *CRAN Task View: Extreme Value Analysis*. URL: `https://CRAN.R-project.org/view=ExtremeValue`, last accessed on 03/03/2017.

Einmahl, J. H. J., Fils-Villetard, A. and Guillou, A. (2008). Statistics of Extremes Under Random Censoring. *Bernoulli*, **14**(1), 207–227.

Einmahl, J. H. J. and Magnus, J. R. (2008). Records in Athletics Through Extreme-Value Theory. *J. Amer. Statist. Assoc.*, **103**(484), 1382–1391.

Einmahl, J. H. J., Zhou, C. and de Haan, L. (2016). Statistics of Heteroscedastic Extremes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **78**(1), 31–51.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin Heidelberg.

Engelen, S., Hubert, M. and Vanden Branden, K. (2005). A Comparison of Three Procedures for Robust PCA in High Dimensions. *Austrian J. Stat.*, **34**(2), 117–126.

European Seismological Commission (1998). *European Macroseismic Scale 1998: EMS-98*. URL: `http://media.gfz-potsdam.de/gfz/sec26/resources/documents/PDF/EMS-98_Original_englisch.pdf`, last accessed on 21/02/2017.

Fackler, M. (2013). Reinventing Pareto: Fits for Both Small and Large Losses. In: ASTIN Colloquium, Den Haag.

Falk, M. and Guillou, A. (2008). Peaks-over-Threshold Stability of Multivariate Generalized Pareto Distributions. *J. Multivariate Anal.*, **99**(4), 715–734.

Fay, M. P. and Shaw, P. A. (2010). Exact and Asymptotic Weighted Logrank Tests for Interval Censored Data: The interval `R` Package. *J. Stat. Softw.*, **36**(2), 1–34.

Ferreira, A., de Haan, L. and Peng, L. (2003). On Optimising the Estimation of High Quantiles of a Probability Distribution. *Statistics*, **37**(5), 401–434.

Ferro, C. A. T. and Segers, J. (2003). Inference for Clusters of Extreme Values. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **65**(2), 545–556.

Filzmoser, P., Fritz, H. and Kalcher, K. (2014). *pcaPP: Robust PCA by Projection Pursuit*. URL: `http://CRAN.R-project.org/package=pcaPP`, `R` package version 1.9-50.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting Forms of the Frequency Distribution of the Largest and Smallest Member of a Sample. *Proc. Cambridge Phil. Soc.*, **24**(2), 180–190.

Fraga Alves, I. and Neves, C. (2014). Estimation of the Finite Right Endpoint in the Gumbel domain. *Statist. Sinica*, **24**(4), 1811–1835.

Fraga Alves, I., Neves, C. and Rosário, P. (2017). A General Estimator for the Right Endpoint With an Application to Supercentenarian Women's Records. *Extremes*, **20**(1), 199–237.

Fréchet, M. (1927). Sur la loi de Probabilité de l'écart Maximum. *Ann. Soc. Math. Polon.*, **6**(3), 93–116. In French.

Frees, E. W. and Valdez, E. A. (1998). Understanding Relationships Using Copulas. *N. Am. Actuar. J.*, **2**(1), 1–25.

Gibowicz, S. and Kijko, A. (1994). *An Introduction to Mining Seismology*. Academic Press, San Diego, CA.

Gil Bellosta, C. J. (2011). *ADGofTest: Anderson-Darling GoF test*. URL: `https://CRAN.R-project.org/package=ADGofTest`, `R` package version 0.3.

Gnedenko, B. V. (1943). Sur la Distribution Limite du Terme Maximum d'une Série Aléatoire. *Ann. of Math.*, **44**(3), 423–453. In French.

Greco, L. and Farcomeni, A. (2016). A Plug-in Approach to Sparse and Robust Principal Component Analysis. *TEST*, **25**(3), 449–481.

Gumbel, E. J. (1958). *Statistics of Extremes*. Columbia University Press, New York, NY.

Gutenberg, B. and Richter, C. F. (1956). Earthquake Magnitude, Intensity, Energy and Acceleration. *Bull. Seismol. Soc. Am.*, **46**(2), 105–145.

Hall, P. (1982). On Some Simple Estimates of an Exponent of Regular Variation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **44**(1), 37–42.

Hampel, F. R. (1971). A General Qualitative Definition of Robustness. *Ann. Math. Statist.*, **42**(6), 1887–1896.

Hampel, F. R. (1974). The Influence Curve and its Role in Robust Estimation. *J. Amer. Stat. Assoc.*, **69**(346), 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics. The Approach Based on Influence Functions*. John Wiley & Sons, Inc., New York, NY.

Hill, B. M. (1975). A Simple General Approach to Inference About the Tail of a Distribution. *Ann. Statist.*, **3**(5), 1163–1174.

Holschneider, M. and Zöller, G. (2014). Induced Seismicity: What is the Size of the Largest Expected Earthquake? *Bull. Seismol. Soc. Am.*, **104**(6), 3153–3158.

Holschneider, M., Zöller, G., Clements, R. and Schorlemmer, D. (2014). Can we Test for the Maximum Possible Earthquake Magnitude? *J. Geophys. Res. Solid Earth*, **119**(3), 2019–2028.

Holschneider, M., Zöller, G. and Hainzl, S. (2011). Estimation of the Maximum Possible Magnitude in the Framework of the Doubly Truncated Gutenberg–Richter Model. *Bull. Seismol. Soc. Am.*, **101**(4), 1649–1659.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables Into Principal Components. *J. Educ. Psychol.*, **24**(6), 417–441.

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, **28**(3/4), 321–377.

Hsing, T. (1991). On Tail Index Estimation Using Dependent Data. *Ann. Statist.*, **19**(3), 1547–1569.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *Ann. Math. Statist.*, **35**(1), 73–101.

Huber, P. J. (1967). The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions. In: Proc. Fifth Berkeley Symp. on Math. Statist. and Prob. (Univ. of Calif. Press), 221–233.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons, Inc., New York, NY.

Hubert, M., Dierckx, G. and Vanpaemel, D. (2013). Detecting Influential Data Points for the Hill Estimator in Pareto-Type Distributions. *Comput. Statist. Data Anal.*, **66**, 13–28.

Hubert, M., Reynkens, T., Schmitt, E. and Verdonck, T. (2016). Sparse PCA for High-Dimensional Data With Outliers. *Technometrics*, **58**(4), 424–434.

Hubert, M., Rousseeuw, P. J. and Vanden Branden, K. (2005). ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, **47**(1), 64–79.

Hubert, M., Rousseeuw, P. J. and Verboven, S. (2002). A Fast Method for Robust Principal Components With Applications to Chemometrics. *Chemometr. Intell. Lab. Syst.*, **60**(1-2), 101–111.

Hubert, M., Rousseeuw, P. J. and Verdonck, T. (2009). Robust PCA for Skewed Data and Its Outlier Map. *Comput. Statist. Data Anal.*, **53**(6), 2264–2274.

Hubert, M. and Vanden Branden, K. (2003). Robust Methods for Partial Least Squares Regression. *J. Chemometr.*, **17**(10), 537–549.

Hubert, M. and Vandervieren, E. (2008). An Adjusted Boxplot for Skewed Distributions. *Comput. Statist. Data Anal.*, **52**(12), 5186–5201.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer-Verlag, New York, NY. 2nd edition.

Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003). A Modified Principal Component Technique Based on the LASSO. *J. Comput. Graph. Statist.*, **12**(3), 531–547.

Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation From Incomplete Observations. *J. Amer. Statist. Assoc.*, **53**(282), 457–481.

Kibler, D. F., Aha, D. W. and Albert, M. K. (1989). Instance-Based Prediction of Real-valued Attributes. *Comput. Intell.*, **5**(2), 51–57.

Kijko, A. (2012). On Bayesian Procedure for Maximum Earthquake Magnitude Estimation. *Res. Geophys.*, **2**(1), 46–51.

Kijko, A., Lasocki, S. and Graham, G. (2001). Non-parametric Seismic Hazard in Mines. *Pure Appl. Geophys.*, **158**(9), 1655–1675.

Kijko, A. and Sellevoll, M. (1989). Estimation of Earthquake Hazard Parameters From Incomplete Data Files. Part I. Utilization of Extreme and Complete Catalogs With Different Threshold Magnitudes. *Bull. Seism. Soc. Am.*, **79**(3), 645–654.

Kijko, A. and Singh, M. (2011). Statistical Tools for Maximum Possible Earthquake Estimation. *Acta Geophys.*, **59**(4), 674–700.

Kijko, A., Smit, A. and Van De Coolwijk, N. (2015). A Scenario Approach to Estimate the Maximum Foreseeable Loss for Buildings due to an Earthquake in Cape Town. *S. Afr. Actuar. J.*, **15**(1), 1–30.

Kiriliouk, A., Rootzén, H., Segers, J. and Wadsworth, J. L. (2016). Peaks Over Thresholds Modelling With Multivariate Generalized Pareto Distributions, available on arXiv:1612.01773.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York, NY. 2nd edition.

Klugman, S. A., Panjer, H. H. and Willmot, G. E. (2012). *Loss Models: From Data to Decisions*. John Wiley & Sons, Inc., Hoboken, NJ. 4th edition.

Klugman, S. A., Panjer, H. H. and Willmot, G. E. (2013). *Loss Models: Further Topics*. John Wiley & Sons, Inc., Hoboken, NJ.

Lee, D., Li, W. K. and Wong, T. S. T. (2012). Modeling Insurance Claims via a Mixture Exponential Model Combined With Peaks-Over-Threshold Approach. *Insurance Math. Econom.*, **51**(3), 538–550.

Lee, S. C. K. and Lin, X. S. (2010). Modeling and Evaluating Insurance Losses via Mixtures of Erlang Distributions. *N. Am. Actuar. J.*, **14**(1), 107–130.

Lee, S. C. K. and Lin, X. S. (2012). Modeling Dependent Risks With Multivariate Erlang mixtures. *Astin Bull.*, **42**(1), 153–180.

Lemberge, P., De Raedt, I., Janssens, K. H., Wei, F. and Van Espen, P. J. (2000). Quantitative Z-Analysis of the $16-17$th Century Archaelogical Glass Vessels Using PLS Regression of EPXMA and μ-XRF Data. *J. Chemometr.*, **14**(5-6), 751–763.

Li, G. and Chen, Z. (1985). Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo. *J. Amer. Statist. Assoc.*, **80**(391), 759–766.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T. and Cohen, K. L. (1999). Robust Principal Component Analysis for Functional Data. *TEST*, **8**(1), 1–73.

Maronna, R. (2005). Principal Components and Orthogonal Regression Based on Robust Scales. *Technometrics*, **47**(3), 264–273.

Maronna, R. A., Martin, D. R. and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. John Wiley & Sons, Ltd., Chichester.

Massart, P. (1990). The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *Ann. Probab.*, **18**(3), 1269–1283.

McLachlan, G. and Peel, D. (2001). *Finite Mixture Models*. John Wiley & Sons, Inc., Hoboken, NJ.

McNeil, A. J. (1997). Estimating the Tails of Loss Severity Distributions Using Extreme Value Theory. *Astin Bull.*, **27**(1), 117–137.

McNeil, A. J., Frey, R. and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, NJ.

Nadarajah, S. and Bakar, S. A. A. (2014). New Composite Models for the Danish Fire Insurance Data. *Scand. Actuar. J.*, **2014**(2), 180–187.

NAM (2016). *Groningen Seismic Hazard and Risk Assessment: Report on Mmax Expert Workshop*. URL: `http://feitenencijfers.namplatform.nl/download/rapport/cef44262-323a-4a34-afa8-24a5afa521d5?open=true`, last accessed on 21/02/2017.

Neuts, M. F. (1981). *Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach*. John Hopkins University Press, Baltimore, MD.

Page, R. (1968). Aftershocks and Microaftershocks of the Great Alaska Earthquake of 1964. *Bull. Seismol. Soc. Am.*, **58**(3), 1131–1168.

Panjer, H. H. (2006). *Operational Risk: Modeling Analytics*. John Wiley & Sons, Inc., Hoboken, NJ.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philos. Mag.*, **11**(2), 559–572.

Pelata, M., Giannopoulos, P. and Haworth, H. (2012). *PCA Unleashed*. URL: `https://research-and-analytics.csfb.com/docView?docid=GaEE3h`, Credit Suisse research report; last accessed on 24/02/2017.

Peters, G. W. and Shevchenko, P. V. (2015). *Advances in Heavy Tailed Risk Modeling: A Handbook of Operational Risk*. John Wiley & Sons, Inc., Hoboken, NJ.

Pfaff, B. and McNeil, A. (2012). *evir: Extreme Values in R*. URL: `https://CRAN.R-project.org/package=evir`, `R` package version 1.7-3.

Pfaff, B. and McNeil, A. (2016). *QRM: Provides R-Language Code to Examine Quantitative Risk Management Concepts*. URL: `https://CRAN.R-project.org/package=QRM`, `R` package version 0.4-13.

Pickands III, J. (1975). Statistical Inference Using Extreme Order Statistics. *Ann. Statist.*, **3**(1), 119–131.

Pigeon, M. and Denuit, M. (2011). Composite Lognormal-Pareto Model With Random Threshold. *Scand. Actuar. J.*, **2011**(3), 177–192.

Pisarenko, V. F., Lyubushin, A. A., Lysenko, V. B. and Golubeva, T. V. (1996). Statistical Estimation of Seismic Hazard Parameters: Maximum Possible Magnitude and Related Parameters. *Bull. Seismol. Soc. Am.*, **86**(3), 691–700.

Plevka, V., Segaert, P., Tampère, C. M. J. and Hubert, M. (2016). Analysis of Travel Activity Determinants Using Robust Statistics. *Transportation*, **43**(6), 979–996.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. URL: http://www.R-project.org/.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. URL: http://www.R-project.org/.

Raschke, M. (2012). Inference for the Truncated Exponential Distribution. *Stoch. Environ. Res. Risk Assess.*, **26**(1), 127–138.

Reiss, R.-D. and Thomas, M. (2007). *Statistical Analysis of Extreme Values, With Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhaüser, Basel. 3rd edition.

Resnick, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer-Verlag, New York, NY.

Reynkens, T. (2017). *rospca: Robust Sparse PCA using the ROSPCA Algorithm*. URL: https://CRAN.R-project.org/package=rospca, R package version 1.0.2.

Reynkens, T. and Verbelen, R. (2017). *ReIns: Functions from "Reinsurance: Actuarial and Statistical Aspects"*. URL: https://CRAN.R-project.org/package=ReIns, R package version 1.0.3.

Reynkens, T., Verbelen, R., Beirlant, J. and Antonio, K. (2017). Modelling Censored Losses Using Splicing: a Global Fit Strategy With Mixed Erlang and Extreme Value Distributions, available on arXiv:1608.01566.

Robson, D. S. and Whitlock, J. H. (1964). Estimation of a Truncation Point. *Biometrika*, **51**(1-2), 33–39.

Rootzén, H., Segers, J. and Wadsworth, J. L. (2016). Multivariate Peaks Over Thresholds Models, available on arXiv:1603.06619.

Rootzén, H. and Tajvidi, N. (2006). Multivariate Generalized Pareto Distributions. *Bernoulli*, **12**(5), 917–930.

Rousseeuw, P. J. (1984). Least Median of Squares Regression. *J. Amer. Statist. Assoc.*, **79**(388), 871–880.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the Median Absolute Deviation. *J. Amer. Statist. Assoc.*, **88**(424), 1273–1283.

Rousseeuw, P. J., Raymaekers, J. and Hubert, M. (2016). A Measure of Directional Outlyingness with Applications to Image Data and Video, available on arXiv:1608.05012.

Ruppert, D. (2010). *Statistics and Data Analysis for Financial Engineering*. Springer, New York, NY.

Rytgaard, M. (1996). Simulation Experiments on the Mean Residual Lifetime Function. In: Proceedings of the XXVII ASTIN Colloquium, Copenhagen, Denmark, 59–81.

Scholz, C. H. (1968). The Frequency-Magnitude Relation of Microfracturing in Rock and its Relation to Earthquakes. *Bull. Seismol. Soc. Am.*, **58**(1), 399–415.

Schwarz, G. E. (1978). Estimating the Dimension of a Model. *J. Amer. Statist. Assoc.*, **6**(2), 461–464.

Scollnik, D. P. M. (2007). On Composite Lognormal-Pareto Models. *Scand. Actuar. J.*, **2007**(1), 20–33.

Scollnik, D. P. M. and Sun, C. (2012). Modeling With Weibull-Pareto Models. *N. Am. Actuar. J.*, **16**(2), 260–272.

Segaert, P., Hubert, M. and Rousseeuw, P. (2017). *mrfDepth: Depth Measures in Multivariate, Regression and Functional Settings*. URL: `https://CRAN.R-project.org/package=mrfDepth`, `R` package version 1.0.3.

Sintubin, M. (2016). *De Mmax van Groningen*. URL: `https://earthlymattersblog.wordpress.com/2016/08/12/de-mmax-van-groningen/`, in Dutch; last accessed on 27/02/2017.

Sun, P. and Zhou, C. (2014). Diagnosing the Distribution of GARCH Innovations. *J. Empir. Financ.*, **29**, 287–303.

Teodorescu, S. and Panaitescu, E. (2009). On the Truncated Composite Weibull-Pareto Model. *Math. Rep. (Bucur.)*, **11**(61), 259–273.

Tijms, H. C. (1994). *Stochastic Models: an Algorithmic Approach*. John Wiley & Sons, Ltd, Chichester, UK.

Todorov, V. (2014). *rrcovHD: Robust Multivariate Methods for High Dimensional Data*. URL: `http://CRAN.R-project.org/package=rrcovHD`, `R` package version 0.2-3.

Todorov, V. and Filzmoser, P. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. *J. Stat. Softw.*, **32**(3), 1–47.

Todorov, V. and Filzmoser, P. (2013). Comparing Classical and Robust Sparse PCA. In: R. Kruse, M. R. Berthold, C. Moewes, M. Á. Gil, P. Grzegorzewski and O. Hryniewicz (eds.), *Synergies of Soft Computing and Statistics for Intelligent Data Analysis*, Springer-Verlag, Berlin Heidelberg, pp. 283–291.

Tukey, J. W. (1960). A Survey of Sampling From Contaminated Distributions. In: I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann (eds.), *Contributions to Probability and Statistics*, Stanford University Press, Stanford, CA, pp. 448–485.

Turnbull, B. W. (1976). The Empirical Distribution Function With Arbitrarily Grouped, Censored and Truncated Data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **38**(3), 290–295.

Utsu, T. (1965). A Method for Determining the Value of b in a Formula $\log n = a - bM$ Showing the Magnitude-Frequency Relation for Earthquakes. *Geophys. Bull. Hokkaido Univ.*, **13**, 99–103. In Japanese with English summary.

van den Beukel, J. (2016). *Groningen Gas Production and Earthquakes*. URL: `https://jillesonenergy.wordpress.com/2016/12/07/groningen-gas-production-and-earthquakes/`, last accessed on 21/02/2017.

Van der Veeken, S. and Hubert, M. (2008). Outlier Detection for Skewed Data. *J. Chemometr.*, **22**(3-4), 235–246.

van der Voort, N. and Vanclay, F. (2015). Social Impacts of Earthquakes Caused by gas Extraction in the Province of Groningen, The Netherlands. *Environ. Impact Assess. Rev.*, **50**, 1–15.

van Eck, T., Goutbeek, F., Haak, H. and Dost, B. (2006). Seismic Hazard due to Small-Magnitude, Shallow-Source, Induced Earthquakes in The Netherlands. *Eng. Geol.*, **87**(1-2), 105–121.

Vanden Branden, K. and Hubert, M. (2005). Robust Classification in High Dimensions Based on the SIMCA Method. *Chemometr. Intell. Lab. Syst.*, **79**(1-2), 10–21.

Vandewalle, B., Beirlant, J., Christmann, A. and Hubert, M. (2007). A Robust Estimator for the Tail Index of Pareto-Type Distributions. *Comput. Statist. Data Anal.*, **51**(12), 6252–6268.

Verbelen, R., Antonio, K. and Claeskens, G. (2016). Multivariate Mixtures of Erlangs for Density Estimation Under Censoring. *Lifetime Data Anal.*, **22**(3), 429–455.

Verbelen, R., Gong, L., Antonio, K., Badescu, A. and Lin, S. (2015). Fitting Mixtures of Erlangs to Censored and Truncated Data Using the EM Algorithm. *Astin Bull.*, **45**(3), 729–758.

Verster, A., de Waal, D., Schall, R. and Prins, C. (2012). A Truncated Pareto Model to Estimate the Under Recovery of Large Diamonds. *Math. Geosci.*, **44**(1), 91–100.

von Mises, R. (1936). La Distribution de la Plus Grande de n Valeurs. *Rev. Math. Union Interbalcanique*, **1**, 141–160. In French.

Weissman, I. (1978). Estimation of Parameters and Large Quantiles Based on the k Largest Observations. *J. Amer. Statist. Assoc.*, **73**(364), 812–815.

Willems, P. (2009). A Time Series Tool to Support the Multi-Criteria Performance Evaluation of Rainfall-Runoff Models. *Environ. Model. Softw.*, **24**(3), 311–321.

Willmot, G. E. and Lin, X. S. (2011). Risk Modelling With the Mixed Erlang Distribution. *Appl. Stoch. Models Bus. Ind.*, **27**(1), 2–16.

Willmot, G. E. and Woo, J.-K. (2007). On the Class of Erlang Mixtures With Risk Theoretic Applications. *N. Am. Actuar. J.*, **11**(2), 99–115.

Willmot, G. E. and Woo, J.-K. (2015). On Some Properties of a Class of Multivariate Erlang Mixtures With Insurance Applications. *ASTIN Bull.*, **45**(1), 151–173.

Worms, J. and Worms, R. (2014). New Estimators of the Extreme Value Index Under Random Right Censoring, for Heavy-Tailed Distributions. *Extremes*, **17**(2), 337–358.

Würtz, D. and Rmetrics Association (2013). *fExtremes: Rmetrics - Extreme Financial Market Data*. URL: `https://CRAN.R-project.org/package=fExtremes`, `R` package version 3010.81.

Zhou, Z., Li, X., Wright, J., Candès, E. J. and Ma, Y. (2010). Stable Principal Component Pursuit. *IEEE Int. Symp. Info.*, 1518–1522.

Zöller, G. and Holschneider, M. (2016a). The Earthquake History in a Fault Zone Tells us Almost Nothing About $m_{\max}$. *Seismol. Res. Lett.*, **87**(1), 132–137.

Zöller, G. and Holschneider, M. (2016b). The Maximum Possible and the Maximum Expected Earthquake Magnitude for Production-Induced Earthquakes at the Gas Field in Groningen, The Netherlands. *Bull. Seismol. Soc. Am.*, **106**(6), 2917–2921.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse Principal Component Analysis. *J. Comput. Graph. Statist.*, **15**(2), 265–286.

List of publications

Articles

Beirlant, J., Fraga Alves, I. and Reynkens, T. (2017). Fitting Tails Affected by Truncation. *Electron. J. Stat.*, **11**(1), 2026–2065.

Beirlant, J., Schoutens, W., De Spiegeleer, J., Reynkens, T. and Herrmann, K. (2016). Hunting for Black Swans in the European Banking Sector Using Extreme Value Analysis. In: J. Kallsen and A. Papapantoleon (eds.), *Advanced Modelling in Mathematical Finance: In Honour of Ernst Eberlein*, Springer International Publishing, Switzerland, pp. 147–166.

Hubert, M., Reynkens, T., Schmitt, E. and Verdonck, T. (2016). Sparse PCA for High-Dimensional Data With Outliers. *Technometrics*, **58**(4), 424–434.

Reynkens, T., Verbelen, R., Beirlant, J. and Antonio, K. (2017). Modelling Censored Losses Using Splicing: a Global Fit Strategy With Mixed Erlang and Extreme Value Distributions, available on arXiv:1608.01566.

`R` packages

Reynkens, T. (2017). *rospca: Robust Sparse PCA using the ROSPCA Algorithm.* URL: `https://CRAN.R-project.org/package=rospca`, `R` package version 1.0.2.

Reynkens, T. and Verbelen, R. (2017). *ReIns: Functions from "Reinsurance: Actuarial and Statistical Aspects".* URL: `https://CRAN.R-project.org/package=ReIns`, `R` package version 1.0.3.

FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICS
SECTION OF STATISTICS
Celestijnenlaan 200B
B-3001 Leuven
tom.reynkens@kuleuven.be